THE VAGUELETTE-WAVELET DECOMPOSITION APPROACH TO STATISTICAL INVERSE PROBLEMS

Felix Abramovich **Bernard W. Silverman**

School of Mathematics, University of Bristol, Bristol BS8 1TH, U.K.

Abstract

A wide variety of scientific settings have to do with *indirect* noisy measurements. We are interesting in some object $f(t)$ but the data is accessible only about some transform $(Kf)(t)$, where K is some linear operator, and $(Kf)(t)$ is in addition corrupted by noise. Recovering $f(t)$ from indirect observations, one faces a Statistical Linear Inverse Problem (SLIP). The usual linear methods for SLIPs, like windowed SVD, do not perform satisfactorily when the original function $f(t)$ is "spatially inhomogeneous". As an alternative, Donoho (1995) suggested the use of wavelet techniques for SLIPs and proposed a method, named wavelet-vaguelette decomposition (WVD), based on the expansion of the unknown $f(t)$ in wavelet series.

In this paper we consider an alternative vaguelette-wavelet decomposition (VWD) which is also based on wavelet expansion. However, in contrast to wavelet-vaguelette decomposition, we expand the observed signal in wavelet series. We discuss theoretical properties of vaguelette-wavelet decomposition and compare its performance with wavelet-vaguelette decomposition and windowed SVD in the context of numerical differentiation of several test functions.

Some key words: Exact risk analysis; Near-minimax estimation; Singular Value Decomposition; Spatially adaptive estimation; Statistical linear inverse problems; Vaguelettes; Wavelets.

1 Introduction.

In a wide variety of statistical problems the direct observations on the object of interest are inaccessible and one faces the problem of estimation and inference from indirect experimental data.

Suppose we wish to estimate an unknown function $f(t)$ but we can observe data only about $(Kf)(t)$, where K is some linear operator. Moreover, the data are corrupted by noise, so that the observed data $y(t)$ are

$$y(t) = (Kf)(t) + \varepsilon(t), \tag{1}$$

where $\varepsilon(t)$ is a white noise with variance σ^2. The problem of recovering f from noisy data y in model (1) is called a Statistical Linear Inverse Problem (SLIP). For most SLIPs in practice one cannot recover f simply as $\hat{f} = K^{-1}y$ because the inverse operator K^{-1} either does not exist at all, or is an unbounded operator, which means that small changes in y would cause large changes in $K^{-1}y$. Such SLIPs are called ill-posed, and are usually treated by applying some regularization procedure (see Tikhonov & Arsenin, 1977 for general theory and O'Sullivan, 1986 (with discussion) for applying in SLIPs) often in combination with Singular Value Decomposition (SVD) methods (see, for example, Johnstone & Silverman, 1990, 1991).

Donoho (1995) proposed the wavelet-vaguelette decomposition (WVD) which works by expanding the function f in a wavelet series, constructing a corresponding vaguelette series for Kf, and then estimating the coefficients using a suitable thresholding approach.

As an alternative we propose the use of vaguelette-wavelet decomposition (VWD) where Kf is expanded in a wavelet series. The corresponding wavelet coefficients of Kf are estimated by thresholding empirical wavelet coefficients of the data. Mapping them back into the vaguelette expansion in the original space yields the VWD-estimator of f.

VWD has some important computational advantages over WVD. In addition, comparing WVD and VWD by numerical differentiation of several test functions, we show that, although the two methods are comparable, if anything VWD slightly outperforms WVD in the examples we consider.

The paper is set out as follows. In Section 2 we review the SVD and WVD approaches, providing the framework within which the VWD method is then defined and discussed. In Section 3, exact risk formulae are obtained for the various approaches, and these are used in Section 4 to carry out a comparison on several examples without any need for simulation. Finally, in Section 5, some theoretical minimax properties of estimators are explored.

2 Different approaches to SLIPs.

2.1 Singular value decompositions and Fourier series.

The underlying idea of SVD methods is the use of a pseudo-inverse operator $(K^*K)^{-1}K^*$, where K^* is the adjoint operator. The unknown f is expanded in a series of the eigenfunctions of the self-adjoint operator K^*K as

$$f = \sum_j \gamma_j^{-1}[Kf, h_j]e_j, \tag{2}$$

where (γ_j^2) and (e_j) are, respectively, the eigenvalues and eigenfunctions of K^*K. The normalized images h_j are defined by $h_j = K(e_j)/||K(e_j)||$ and $[\cdot,\cdot]$ is the standard inner product in L_2. In the presence of noise in the data we replace Kf in (2) by y and define the estimator $\hat{f}$ of f as

$$\hat{f}^{SVD} = \sum_j \gamma_j^{-1}[y, h_j]e_j \tag{3}$$

However, the ill-posed problems are characterized by the fact that the corresponding eigenvalues of K^*K tend to zero. To prevent division by near-zero eigenvalues γ_j's in (3) one can define a windowed SVD-estimator:

$$\hat{f}_w^{SVD} = \sum_j w_j\gamma_j^{-1}[y, h_j]e_j,$$

where the weights w_j's are chosen so that $\sum w_j^2 \gamma_j^{-2} < \infty$. Choosing different weights essentially leads to some other procedures for SLIPs as special cases. Thus, for example, setting $w_j = \gamma_j^2/(\gamma_j^2 + \lambda)$, where λ is a regularization parameter, gives a regularization method of Tikhonov & Arsenin (1977).

SVD has been widely used in practice in various theoretical and applied problems. Moreover, Johnstone & Silverman (1990, 1991) showed that the windowed SVD-estimator with proper chosen windows is asymptotically the best estimator in a certain minimax sense.

On the other hand, despite all its attractive properties SVD faces certain problems. Applying windowed SVD we, essentially, assume that f has a parsimonious representation in the e_j's, that is only few first coefficients in (3) are "significant" and represent a real signal. However, this basis is defined entirely by the operator K and ignores the specific physical nature of the problem under study. It may suit certain classes of functions, for example, those considered in the theoretical framework of Johnstone & Silverman (1990, 1991), but be inappropriate for others. This feature causes limitations in using SVD. For example, for stationary operators, such as integration and differentiation, the corresponding eigenfunctions generate Fourier sine and cosine basis. Fourier series are appropriate for representation of spatially homogeneous functions, but fail in approximating inhomogeneous signals which are smooth in some regions while have rapid local changes in others. In order to approximate the oscillations one has to use high frequency components of Fourier series that badly contaminate the estimate in smooth regions. Thus, for such operators the use of SVD inherently restricts one within the class of spatially homogeneous functions.

2.2 The wavelet-vaguelette decomposition.

In response to these limitations of SVD for solving SLIPs, Donoho (1995) proposed the so-called Wavelet-Vaguelette Decomposition (WVD) algorithm which is based on expanding the function f in wavelet series. Wavelet series are generated by translations and dilations of a single fixed function ψ, called the mother wavelet: $\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$, $j, k \in Z$. The examples of mother wavelets can be found in Chiu (1992), Daubechies (1992). Wavelets can be easily calculated and, even more important, they are localized in both the time and frequency domains. This characteristic property allows parsimonious wavelet representation of a wide class of functions. Choosing the mother wavelet with corresponding regularity properties one can generate not only orthonormal bases in global $L_2(R^d)$ but unconditional bases in a wide range of more specific spaces. This set includes, in particular, the well-known Sobolev and Hölder scales of smooth functions, but in addition less "traditional" spaces such as functions of bounded variation. See Meyer (1992), Donoho & Johnstone (1992), Donoho et al. (1995) for rigorous definitions and details.

In the WVD approach, the unknown function f is represented in terms of a wavelet expansion

$$f = \sum_j \sum_k [f, \psi_{jk}]\psi_{jk}$$

Let $\Psi_{jk} = K\psi_{jk}$. The crucial point of WVD decomposition is that for some operators K there exist constants β_{jk} such that the set of scaled functions $v_{jk} = \Psi_{jk}/\beta_{jk}$ form a Riesz basis in $R(K)$ in L_2-norm, that is there exist two constants $0 < A \leq B < \infty$ such that

$$A \sum_j \sum_k c_{jk}^2 \leq || \sum c_{jk} v_{jk} ||^2 \leq B \sum_j \sum_k c_{jk}^2 \tag{4}$$

for all square-summable sequences $\{c_{jk}\}$. The v_{jk}'s are called vaguelettes.

Obviously, only special operators K imply (4). However, (4) holds, for example, for homogeneous operators satisfying $K[f(at)] = a^\alpha K[f(t)]$ for some α. Examples of such operators include differentiation, integration, fractional integration and Radon transform. For homogeneous operators $\beta_{jk} = 2^{-j\alpha}$ (Donoho, 1995) and the corresponding vaguelettes v_{jk}'s are translations and dilations of a single mother function and are, therefore, "almost wavelets" but without orthogonality. The property (4) also holds for various convolution operators; see Donoho (1995). In what follows we consider only operators K satisfying (4).

Provided the wavelet basis ψ_{jk} is chosen appropriately, any function $g \in R(K)$ can be expanded in a vaguelette series as

$$g = \sum_j \sum_k [g, u_{jk}] v_{jk},$$

where $\{u_{jk}\}$ is a dual vaguelette basis satisfying $K^* u_{jk} = \beta_{jk} \psi_{jk}$. Dual bases $\{u_{jk}\}$ and $\{v_{jk}\}$ are biorthogonal, i.e. $[v_{jk}, u_{lm}] = \delta_{jl} \delta_{km}$.

If we observed the noiseless signal $Kf \in R(K)$ we could expand it in a vaguelette series:

$$Kf = \sum_j \sum_k [Kf, u_{jk}] v_{jk}$$

and then recover the original function f as

$$f = \sum_j \sum_k [Kf, u_{jk}] \beta_{jk}^{-1} \psi_{jk} = \sum_j \sum_k [Kf, \tilde{\Psi}_{jk}] \psi_{jk}, \tag{5}$$

where $\tilde{\Psi}_{jk} = u_{jk}/\beta_{jk}$ and, hence, $K^* \tilde{\Psi}_{jk} = \psi_{jk}$. The above formula (5) is a key formula for WVD.

In the case of noisy data, we expand the observed signal y in terms of vaguelettes, with coefficients $\hat{b}_{jk} = [y, \tilde{\Psi}_{jk}]$ which satisfy

$$\hat{b}_{jk} = b_{jk} + w_{jk}, \tag{6}$$

where $b_{jk} = [Kf, \tilde{\Psi}_{jk}]$ are the noiseless vaguelette coefficients from (5) and $w_{jk} = [\epsilon, \tilde{\Psi}_{jk}]$ are the vaguelette decomposition of a white noise.

From (6), $\hat{b}_{jk} \sim \mathcal{N}(b_{jk}, \sigma^2 ||\tilde{\Psi}_{jk}||^2)$. Construct rescaled coefficients $\hat{b}^0_{jk} = \hat{b}_{jk}/||\tilde{\Psi}_{jk}||$ which will all have the same variance σ^2. Extraction of the significant $\hat{b}^0_{jk}$'s is then based on the idea that only "large" $|\hat{b}^0_{jk}|$'s contribute a real signal and can be naturally performed by thresholding: $\delta_\lambda(\hat{b}^0_{jk}) = \mathrm{sign}(\hat{b}^0_{jk})(|\hat{b}^0_{jk}| - \lambda)_+$ (soft thresholding) or $\delta_\lambda(\hat{b}^0_{jk}) = \hat{b}^0_{jk} 1\{|\hat{b}^0_{jk}| > \lambda\}$ (hard thresholding) for some threshold value $\lambda \geq 0$. It is important to note that the linear weighting of eigenvalues in windowed SVD is replaced here by the nonlinear thresholding of vaguelette coefficients and, hence, the corresponding WVD-estimator is nonlinear. Mapping the thresholded coefficients back into wavelet expansion in the original space yields the resulting WVD-estimator $\hat{f}_\lambda^{WVD}$:

$$\hat{f}_\lambda^{WVD} = \sum_j \sum_k ||\tilde{\Psi}_{jk}|| \delta_\lambda(\hat{b}_{jk}^0) \psi_{jk} \tag{7}$$

In particular, for homogeneous operators the $\tilde{\Psi}_{jk}$'s will be translations and dilations of a single mother vaguelette $\tilde{\Psi}_{00}$: $\tilde{\Psi}_{jk} = 2^{j\alpha} 2^{j/2} \tilde{\Psi}_{00}(2^j t - k)$. This makes all computations substantially easier. The norms $||\Psi_{jk}||$'s are equal at each level j and, therefore, although the variances of the vaguelette coefficients on the j-th level increase like $2^{2j\alpha}$ as a direct consequence of the ill-posedness of the inverse problem, within each given level j all the $\hat{b}_{jk}$'s have the same variance. Using this fact, Donoho (1995) does not normalize $\tilde{\Psi}_{jk}$'s but suggests instead level-dependent thresholds proportional to $2^{j\alpha}$. Obviously, this is equivalent to global thresholding of normalized coefficients and Donoho's WVD-estimator coincides with (7). Kolaczyk (1994) considers in details the WVD algorithms for SLIPs with the Radon transform operator which arises, for example, in positron emission tomography.

Donoho (1995) showed that the WVD-estimator, provided with optimally chosen threshold λ, is an asymptotically near-minimax estimator within a wide range of various spaces including those containing spatially inhomogeneous functions (see details in Section 5 below). Moreover, he proved that for such functions WVD outperforms any linear estimator and, hence, windowed SVD in particular.

2.3 The vaguelette-wavelet decomposition.

A natural alternative to WVD is the Vaguelette-Wavelet Decomposition (VWD): expand the observed data y in terms of wavelets, threshold appropriately the corresponding wavelet coefficients, and then map back by K^{-1} to obtain an estimate of f in terms of vaguelette series. Thus, it is Kf, rather than f, that is expanded in a wavelet series.

Suppose we have a wavelet expansion

$$Kf = \sum_j \sum_k d_{jk} \psi_{jk}, \tag{8}$$

where ψ_{jk}'s are wavelets constructed to ensure that $\psi_{jk} \in R(K)$ for all j and k, and $d_{jk} = [Kf, \psi_{jk}]$. We should note that, although for convenience we keep the same notation for wavelets, the ψ_{jk}'s are now wavelets in $R(K)$ and generally will be different from those in the original domain $D(K)$ used in Section 2.2. The same will be true for vaguelettes introduced below. Assume the existence of constants β_{jk}'s, different from those used in Section 2.2, such that (4) holds for $v_{jk} = K^{-1}\psi_{jk}/\beta_{jk}$. The function f is then recovered from (8) by expanding in the vaguelette series

$$f = \sum_j \sum_k [Kf, \psi_{jk}] \beta_{jk} v_{jk} = \sum_j \sum_k [Kf, \psi_{jk}] \Psi_{jk}, \tag{9}$$

where $\Psi_{jk} = K^{-1}\psi_{jk}$.

As in WVD, the wavelet coefficients of a noisy signal y, $\hat{d}_{jk} = [y, \psi_{jk}]$, are contaminated by noise

$$\hat{d}_{jk} = d_{jk} + w_{jk}, \tag{10}$$

where $w_{jk} = [\varepsilon, \psi_{jk}]$ are the coefficients of the wavelet decomposition of a white noise, and need to be denoised by thresholding. The resulting VWD-estimator $\hat{f}_\lambda^{VWD}$ will then be

$$\hat{f}_\lambda^{VWD} = \sum_j \sum_k \delta_\lambda([y, \psi_{jk}]) \Psi_{jk}, \tag{11}$$

where $\delta_\lambda(\cdot)$ is a soft or hard thresholding operator.

We can of course consider the VWD approach as a "plug-in" estimator, in that we find a good wavelet-based estimator of Kf and then plug it with K^{-1} to estimate f itself. Some theory discussed in Section 5 below can be used to justify this approach.

2.4 The discrete WVD and VWD.

In practice, given the discrete data, we implement both approaches making use of a discrete wavelet or vaguelette transform to find the corresponding empirical wavelet or vaguelette coefficients. For the discrete case, all inner products are re-defined in the l_n^2-sense, $[f,g] = f_1 g_1 + ... + f_n g_n$.

Consider the discrete version of the original model (1)

$$y_i = (Kf)(t_i) + \varepsilon_i,$$

where $t_i = i/n$ and the ε's are independent normal variables $N(0,\sigma^2)$. The Discrete Wavelet Transform (DWT) of the data-vector $y = (y_1, ..., y_n)^T$ derives empirical wavelet coefficients $\hat{d}_{jk}$'s and is given by $\hat{d} = Wy$, where W is the DWT-matrix satisfying $W_{j,i} \approx \psi_{jk}(i/n) = 2^{j/2}\psi(2^j i/n - k)$. The approximate " $\approx$ " is caused by some slight modifications needed to make W orthogonal. The inverse DWT becomes $y = W^T \hat{d}$ due to the orthogonality of W. In practice, if $n = 2^{J+1}$ for some positive J, both DWT and inverse DWT are performed by Mallat's (1989) algorithm which requires only $O(n)$ operations and is available in several standard implementations, for example, in Nason & Silverman's WaveThresh (Nason, 1993; Nason & Silverman, 1994).

By analogy with DWT we can define the (non-orthogonal) Discrete Vaguelette Transform (DVT) of the data y as $b = Vy$, where $V_{jk,i} = \tilde{\Psi}_{jk}(i/n)$. Note that the inverse DVT matrix $V^{-1}_{jk,i} = \Psi_{jk}(i/n)$. The DVT-matrix V varies for different K's, so performing DVT and inverse DVT may be computationally expensive in the general case. Kolaczyk (1994) provides efficient DVT-algorithms for homogeneous operators that require $O(n \log^2 n)$ operations.

3 Derivation of average mean square errors.

A natural measure of the global goodness-of-fit for an estimate $\hat{f}$ of the unknown function f is the mean integrated square error MISE $= \int (f - \hat{f})^2$. In order to compare the performance of the windowed SVD, WVD and VWD approaches, we studied the discrete version of MISE, the average mean square error (AMSE), defined as $\mathrm{AMSE}(\hat{f}; f) = (1/n)E||\hat{f} - f||^2 = (1/n)E\sum(\hat{f}_i - f_i)^2$. In this section we derive exact formulae for the mean square error in individual thresholded vaguelette and wavelet coefficients in (7) and (11) respectively, which are then used subsequently to allow AMSE-comparisons to be made without the need for any simulations.

3.1 MSE for vaguelette and wavelet coefficients.

The DVT of the observed y in WVD followed by appropriate rescaling, implies vaguelette coefficients $\hat{b}^0_{jk} \sim N(b^0_{jk}, \sigma^2)$. Straightforward calculations yield (compare with (46), (47) of Donoho & Johnstone, 1994 obtained for $\sigma = 1$) formulae for the mean square error of the individual vaguelette coefficients, as follows:

$$\begin{aligned} E\left\{\delta_\lambda(\hat{b}^0_{jk}) - b^0_{jk}\right\}^2 &= (b^0_{jk})^2 - \left((b^0_{jk})^2 - \sigma^2 - \lambda^2\right)\left(\tilde{\Phi}\{(\lambda - b^0_{jk})/\sigma\} + \tilde{\Phi}\{(\lambda + b^0_{jk})/\sigma\}\right) \\ &- \sigma(\lambda + b^0_{jk})\phi\{(\lambda - b^0_{jk})/\sigma\} - \sigma(\lambda - b^0_{jk})\phi\{(\lambda + b^0_{jk})/\sigma\} \qquad (12) \end{aligned}$$

for soft thresholding and

$$\begin{aligned} E\left\{\delta_\lambda(\hat{b}^0_{jk}) - b^0_{jk}\right\}^2 &= (b^0_{jk})^2 - \left((b^0_{jk})^2 - \sigma^2\right)\left(\tilde{\Phi}\{(\lambda - b^0_{jk})/\sigma\} + \tilde{\Phi}\{(\lambda + b^0_{jk})/\sigma\}\right) \\ &+ \sigma(\lambda - b^0_{jk})\phi\{(\lambda - b^0_{jk})/\sigma\} + \sigma(\lambda + b^0_{jk})\phi\{(\lambda + b^0_{jk})/\sigma\} \qquad (13) \end{aligned}$$

for hard thresholding, where ϕ, Φ are the standard normal p.d.f. and c.d.f. respectively and $\tilde{\Phi} = 1 - \Phi$.

The DWT of y used in VWD yields noisy wavelet coefficients $\hat{d}_{jk} \sim N(d_{jk}, \sigma^2)$ and by complete analogy with (12) and (13) one immediately has

$$\begin{aligned} E\left\{\delta_\lambda(\hat{d}_{jk}) - d_{jk}\right\}^2 &= d^2_{jk} - \left(d^2_{jk} - \sigma^2 - \lambda^2\right)\left(\tilde{\Phi}\{(\lambda - d_{jk})/\sigma\} + \tilde{\Phi}\{(\lambda + d_{jk})/\sigma\}\right) \\ &- \sigma\,(\lambda + d_{jk})\,\phi\{(\lambda - d_{jk})/\sigma\} - \sigma\,(\lambda - d_{jk})\,\phi\{(\lambda + d_{jk})/\sigma\} \end{aligned}$$

for soft thresholding and

$$
\begin{aligned}
E\left\{\delta_\lambda(\hat{d}_{jk}) - d_{jk}\right\}^2 &= d_{jk}^2 - (d_{jk}^2 - \sigma^2)\left(\tilde{\Phi}\{(\lambda - d_{jk})/\sigma\} + \tilde{\Phi}\{(\lambda + d_{jk})/\sigma\}\right) \\
&+ \sigma\,(\lambda - d_{jk})\,\phi\{(\lambda - d_{jk})/\sigma\} + \sigma\,(\lambda + d_{jk})\,\phi\{(\lambda + d_{jk})/\sigma\}
\end{aligned}
$$

for hard thresholding respectively.

Although the formulae for individual MSE are essentially identical for vaguelette and wavelet coefficients, we should note that the wavelet coefficients $\hat{d}_{jk}$'s are independent beacuse of the orthogonality of DWT, while the vaguelette coefficients $\hat{b}^0_{jk}$'s are not.

3.2 Average mean square error for wavelet-vaguelette decomposition.

Recall that

$$
\hat{f}_\lambda^{WVD} = \sum_j \sum_k ||\tilde{\Psi}_{jk}||\delta_\lambda(\hat{b}^0_{jk})\psi_{jk}
$$

The wavelet basis $\{\psi_{jk}\}$ is orthonormal. Thus, even though the $\hat{b}^0_{jk}$ are not independent, the AMSE of the WVD-estimator is just the sum of their individual MSEs and, therefore,

$$
\mathrm{AMSE}(\hat{f}_\lambda^{WVD}; f) = \frac{1}{n}E||f - \hat{f}_\lambda^{WVD}||^2 = \frac{1}{n}\sum_j \sum_k ||\tilde{\Psi}_{jk}||^2 E\left\{\delta_\lambda(\hat{b}^0_{jk}) - b^0_{jk}\right\}^2, \tag{14}
$$

where $E\left\{\delta_\lambda(\hat{b}^0_{jk}) - b^0_{jk}\right\}^2$ are given by (12) or (13) as appropriate.

3.3 Average mean square error for vaguelette-wavelet decomposition.

The VWD-estimator $\hat{f}_\lambda^{VWD}$ has been defined in (11) as

$$
\hat{f}_\lambda = \sum_j \sum_k \delta_\lambda(\hat{d}_{jk})\Psi_{jk}
$$

Using the fact that the $\hat{d}_{jk}$ are independent, we have

$$\text{AMSE}(\hat{f}_\lambda^{VWD}; f) =$$
$$\frac{1}{n} E \left\{ \sum_j \sum_k \left(\delta_\lambda(\hat{d}_{jk}) - d_{jk} \right) \Psi_{jk} \right\}^2 = \frac{1}{n} \sum_{j,k,i,l} E \left(\delta_\lambda(\hat{d}_{jk}) - d_{jk} \right) \left(\delta_\lambda(\hat{d}_{il}) - d_{il} \right) [\Psi_{jk}, \Psi_{il}] =$$
$$\frac{1}{n} \sum_j \sum_k E \left\{ \delta_\lambda(\hat{d}_{jk}) - d_{jk} \right\}^2 ||\Psi_{jk}||^2 + \frac{1}{n} \sum_{(j \neq i) \vee (k \neq l)} \left(E\delta_\lambda(\hat{d}_{jk}) - d_{jk} \right) \left(E\delta_\lambda(\hat{d}_{il}) - d_{il} \right) [\Psi_{jk}, \Psi_{il}], \tag{15}$$

and by straightforward calculations

$$\begin{aligned} E\delta_\lambda(d_{jk}) - d_{jk} &= d_{jk} \left(\Phi\{(\lambda - d_{jk})/\sigma\} - \Phi\{(-\lambda - d_{jk})/\sigma\} \right) + \sigma \left(\phi\{(\lambda + d_{jk})/\sigma\} - \phi\{(\lambda - d_{jk})/\sigma\} \right) \\ &+ \lambda \left(\Phi\{(\lambda + d_{jk})/\sigma\} - \Phi\{(\lambda - d_{jk})/\sigma\} \right) \end{aligned}$$

or

$$E\delta_\lambda(d_{jk}) - d_{jk} = d_{jk} \left(\Phi\{(\lambda - d_{jk})/\sigma\} - \Phi\{(-\lambda - d_{jk})/\sigma\} \right) + \sigma \left(\phi\{(\lambda + d_{jk})/\sigma\} - \phi\{\lambda - d_{jk})/\sigma\} \right)$$

for soft and hard thresholding rules respectively.

From (14) and (15) we can find ideal optimal thresholds λ that minimize the corresponding AMSEs for particular WVD- and VWD-estimators of a given function f with operator K. Hence, we can compare their performance for different SLIPs and various functions without any simulations. Note that the WVD- and VWD-estimators always have the same AMSE at the two extreme values of λ: AMSE $= (1/n)E||K^{-1}y - f||^2$ for $\lambda = 0$ (no thresholding) and AMSE tends to $(1/n)||f||^2$ as $\lambda \to \infty$ (total thresholding, $\hat{f}_\lambda^{WVD} \equiv \hat{f}_\lambda^{VWD} \equiv 0$).

3.4 Average mean square error for singular value decomposition.

To study the effeciency of the wavelet-based decompositions in solving SLIPs we can also compare their AMSEs with that of windowed SVD as described in Section 2.1.

Recall that, in windowed SVD, f is expanded in terms of the orthogonal eigenfunctions e_j of the operator K^*K. The corresponding estimator $\hat{f}_w^{SVD}$ is

$$\hat{f}_w^{SVD} = \sum_j w_j \hat{c}_j e_j,$$

where $\hat{c}_j = \gamma^{-1}[y, h_j]$ and γ_j, e_j and h_j were defined in (2). Note that $\hat{c}_j \sim N(c_j, \gamma^{-2}\sigma^2)$, where $c_j = \gamma^{-1}[Kf, h_j]$. Due to the orthogonality of the e_j

$$\text{AMSE}(\hat{f}_w^{SVD}; f) = \frac{1}{n}\sum_j \text{MSE}(w_j\hat{c}_j; c_j) = \frac{1}{n}\left(\sum_j (1-w_j)^2 c_j^2 + \sigma^2 \sum_j w_j^2 \gamma_j^{-2}\right). \tag{16}$$

Let $\hat{f}_k^{SVD}$ be the estimator

$$\hat{f}_k^{SVD} = \sum_{j=1}^{k} \hat{c}_j e_j, \tag{17}$$

obtained setting $w_j = 1$ for $j \le k$ and 0 for $j > k$. Then, from (16), we have

$$\text{AMSE}(\hat{f}_k^{SVD}; f) = \frac{1}{n}\left(\sum_{j=k+1} c_j^2 + \sigma^2 \sum_{j=1}^{k} \gamma_j^{-2}\right). \tag{18}$$

For given K and f we can find the optimal number of terms k in the SVD expansion that minimizes the AMSE (18).

4 Comparison between approaches.

In this Section we use AMSE to compare between different estimators. In order to provide a concrete basis on which to compare WVD and VWD, we consider numerical differentiation as a particular example of an ill-posed problem. From (14) and (15) we find ideal optimal WVD- and VWD-estimators that minimize AMSEs for numerical differentiation of various test functions. In addition, we compare them with the optimal windowed SVD-estimator (17).

4.1 WVD and VWD for numerical differentiation.

Suppose that f is an unknown periodic (for simplicity) function on $[0,1]$ and that Kf is the n-vector defined by

$$(Kf)_i = \sum_{j \le i} f(i/n)$$

Suppose we observe a vector $y = (y_1, ..., y_n)^T$ of independent normally distributed observations with means $(Kf)_i$'s and variance σ^2. Define Δ to be the finite difference operator $(\Delta f)_i = f(\frac{i+1}{n}) - f(\frac{i}{n})$.

First we consider WVD. To derive the corresponding vaguelette vectors $\tilde{\Psi}_{jk}$, it is easy to verify that $(K^* \tilde{\Psi}_{jk})_i = \psi_{jk,i}$ implies $\tilde{\Psi}_{jk,i} = -(\Delta \psi_{jk})_i$ and from (7)

$$\hat{f}^{WVD}_{\lambda,i} = -\sum_j \sum_k ||\Delta\psi_{jk}|| \delta_\lambda \left([y, \Delta\psi_{jk}]/||\Delta\psi_{jk}|| \right) \psi_{jk,i} \ .$$

To perform VWD note that $(K^{-1}f)_i = (\Delta f)_{i-1}$ and the corresponding vaguelette vectors are $\Psi_{jk,i} = (K^{-1}\psi_{jk})_i = (\Delta\psi_{jk})_{i-1}$. The resulting VWD-estimator is

$$\hat{f}^{VWD}_{\lambda,i} = \sum_j \sum_k \delta_\lambda([y, \psi_{jk}])(\Delta\psi_{jk})_{i-1},$$

where the wavelet coefficients $[y, \psi_{jk}]$'s are obtained by DWT of the data.

4.2 Examples.

We studied the properties of WVD and VWD estimators in numerical differentiation on three test-functions satisfying periodic boundary conditions on $[0,1]$ (see Fig.1):

1) Bumps (similar to the analogous Donoho & Jonstone, 1994 Bumps example)

$f(t) = \sum_j h_j K((t - t_j)/w_j)$, where $K(t) = \max((1 - |t|)^4, 0)$,

$(t_j) = (0.1, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81)$,

$(h_j) = (4, 5, 3, 4, 5, 4.2, 2.1, 4.3, 3.1, 5.1, 4, 2)$,

$(w_j) = (0.005, 0.005, 0.006, 0.01, 0.01, 0.03, 0.01, 0.01, 0.005, 0.008, 0.005)$

2) Donoho & Johnstone (1994) Blocks example

$f(t) = \sum_j h_j K(t - t_j)$, where $K(t) = (1 + \text{sign}(t))/2$,

(t_j) are the same as in the previous example,

$(h_j) = (4, -5, 3, -4, 5, -4.2, 2.1, 4.3, -3.1, 5.1, -4, 2)$

3) Donoho & Johnstone (1994) HeaviSine example

$f(t) = 4 \sin 4\pi t - \text{sign}(t - 0.3) - \text{sign}(0.72 - t)$

The discrete wavelet or vaguelette transform allows to estimate $n-1$ wavelet or vaguelette coefficients d_{jk} (b_{jk}) for $j = 0, ..., J$, $k = 0, ..., 2^j - 1$. However, one cannot reconstruct a constant function as a finite linear combination of wavelets or vaguelettes even for noiseless data and the n-th degree of freedom is usually used for estimating the overall mean of the function. The additional estimation increases the AMSE. To eliminate its influence and, in addition, to reduce vectors $f = (f(t_1), ..., f(t_n))^T$ to the same l_n^2-norm we studied standardized f_0's defined as $f_0 = (f - \bar{f})/SD(f)$.

We found optimal values of soft thresholds λ's for WVD and VWD estimators that minimize (14) and (15) respectively for each of the three test-cases, at five different noise levels providing signal-to-noise ratio SNR $= SD(Kf_0)/\sigma = 150$, 100, 75, 50 and 25. The number of points n was taken to be 512. Note, that $b_{jk}^0 = [Kf, \tilde{\Psi}_{jk}] = [f, \psi_{jk}]$, so the b_{jk}^0's in (14) can be calculated simply by performing DWT of f.

Since integration increases the order of a function's smoothness by one, the regularity of the mother wavelet in VWD should be larger by one than that in WVD for a proper comparison. We used the compactly supported mother wavelets D_4 and D_6 from Daubechies's family (Daubechies, 1992) in WVD and D_5 and D_7, correspondingly, from the same family in VWD. Table 1 contains AMSEs of WVD- and VWD-estimators for all test cases, signal-to-noise ratios and for the pairs $(D_4,\ D_5)$ and $(D_6,\ D_7)$ of mother wavelets used in WVD and VWD respectively.

To compare WVD and VWD with windowed SVD (WSVD), we calculated the minimal AMSEs for a WSVD-estimator (17) with the optimally chosen k's. The results are also given in Table 1.

All computations were done in the statistical package S-Plus using the WaveThresh software developed by Nason and Silverman (Nason, 1993; Nason & Silverman, 1994).

4.3 Analysis of results.

The analysis of the results summarized in Table 1 leads to several interesting observations.

Table 1 shows that there is no significant difference for WVD and VWD between two pairs of mother wavelet functions. Although there is no strong preference between the wavelet-based methods, for the pair $(D_6,\ D_7)$ VWD uniformly outperforms WVD. It is interesting, that in most cases WVD performed better for the less regular mother wavelet (D_4), while VWD had an opposite tendency.

WVD and VWD indeed outperform WSVD for the Bumps example which is a badly spatially inhomogeneous function. WSVD, however, is preferable in the HeaviSine example. This is probably not so suprising since, apart from two points of discontinuity, the HeaviSine function has a parsimonious Fourier expansion.

More fruitful analysis, however, can be done by studying asymptotic properties of the estimators. Increasing SNR (or decreasing the noise level) is somewhat equivalent to increasing the sample size n and in this sense SNR can be considered as an "surrogate sample size". Thus, we can study the convergence for different methods by analysing their performance as a function of SNR. Table 1 clearly shows that the rates of convergence for WVD and VWD are much faster than that of WSVD. Even for the HeaviSine example, though WSVD is almost twice better than WVD and VWD for SNR $= 25$, the difference rapidly decreases and is negligible for SNR $= 150$. The theoretical ground for this phenomenon is given in Section 5.

5 Theoretical results.

In this Section we consider the theoretical aspects of WVD and VWD more rigorously, establish the asymptotic near-optimality (in the minimax sense) of the VWD-estimator and compare with the analogous results for WVD obtained by Donoho (1995).

First we introduce some definitions and notations. Consider the original model (1) and suppose that the operator K acts on some space $B^s_{p,q}\ (R^d)$ from the Besov scale. This includes, in particular, such well-known spaces as Sobolev's H^s ($B^s_{2,2}$), Hölder's C^s ($B^s_{\infty,\infty}$), spaces of functions of bounded variation ($p = 1, s = 1$) and many other interesting examples. See Meyer, 1992; Donoho & Johnstone, 1992 for rigorous definitions and details. The parameter s measures the number of derivatives whose existence is understood in an L_p-sense, while q provides some further flexibility. The fundamental property of wavelets is that given the mother wavelet of regularity τ, the corresponding wavelet basis is an unconditional basis within the whole range of Besov scale with $s \leq \tau$. This allows a parsimonious wavelet expansion for a wide set of different functions.

Let $D(K) = B^s_{p,q}\ (R^d)$, where $s + 1/2 - 1/p > 0$, and assume the mother wavelet ψ, generating the wavelet basis in $D(K)$, has regularity $\tau \geq s$. Define the risk function $R(\hat{f}; f) = E||\hat{f} - f||^2_{L_2}$ and let R^* be the minimax risk within Besov ball $B^s_{p,q}\ (C_0)$ with radius C_0. We shall call the estimator $\hat{f}$ near-minimax in $B^s_{p,q}\ (C_0)$ if up to a logarithmic factor it acheves the optimal rate of convergence, i.e. if $\sup_{f\in B^s_{p,q}\ (C_0)} R(\hat{f}, f) \leq C \log(n) R^*$ for some constant C.

5.1 Near-optimality of WVD.

The results of Donoho (1995) provide the near-minimax property of WVD. For properly chosen threshold λ^* the corresponding WVD-estimator $f^{WVD}_{\lambda^*}$ is near-minimax within the whole range of Besov balls:

$$R_{WVD} = \sup_{f\in B^s_{p,q}\ (C_0)} R(\hat{f}^{WVD}_{\lambda^*}; f) \leq C_1 \log(n) R^* \tag{19}$$

for all $p, q \geq 1$, for all $C_0 > 0$ and for all $s_0 < s < \tau$.

Thus, even without knowing the exact regularity s of the unknown f and without specifying p, q and C_0, the same WVD estimator is a near-minimax estimator over a wide range of

Besov scale. It will be a near-minimax estimator , for example, over all Sobolev and Hölder classes with $s \leq \tau$ and over various classes of spatially inhomogeneous functions. In this sense the WVD-estimator is spatially adaptive to the unknown smoothness. The logarithmic factor in (19) is a "price" for such spatial adaptivity and cannot be avoided (see also Lepskii, 1990; Goldenshluger & Nemirovski, 1994). Donoho (1995) also showed that for homogeneous operators K satisfying $K[f(at)] = a^{\alpha} K[f(t)]$, both R_{WVD} (up to the logarithmic factor) and R^* tend to zero at the (optimal) rate

$$r = \frac{s}{s + d/p + \alpha},$$

while the rate of convergence for the minimax linear estimator is

$$r' = \frac{s}{s + d/\max(2, p) + \alpha}$$

Thus, no linear estimator (windowed SVD, in particular) can attain the optimal rate within Besov spaces with $p < 2$. Donoho's results confirms our preliminary considerations that WVD should perform better than windowed SVD for spatially inhomogeneous functions.

Note that due to the non-orthogonality of the vaguelette basis, the thresholding of vaguelette coefficients in WVD is performed within coloured noise that increases the near-optimality of the procedure. The "more orthogonal" the vaguelette basis (the closer A and B in (4)), the more optimal is the thresholding and, thus, the less is C_1 in (19).

5.2 Near-optimality of VWD.

To analyse the properties of the VWD procedure we should assume in addition that $K : B^s_{p,q}\ (C_0) \to B^{s'}_{p',q'}\ (C'_0)$, where $s' + 1/2 - 1/p' > 0$, i.e. the image of a Besov ball $B^s_{p,q}\ (C_0) \in D(K)$ is another Besov ball $B^{s'}_{p',q'}\ (C'_0) \in R(K)$. This assumption may seem restrictive but, fortunately, holds for the most commonly used operators defined on Besov spaces, for example, integration or differentiation. Thus, we can expand $Kf \in B^{s'}_{p',q'}$ in terms of wavelets with a mother wavelet of regularity $\tau' > s'$ and perform VWD.

The following theorem establishes the near-optimality (in the minimax sense) of the VWD-procedure:

Theorem 1 *There exists λ^* such that the corresponding VWD-estimator will be near-minimax within all Besov balls $B^s_{p,q}\ (C_0)$ for all $p, q \geq 1$, for all $C_0 > 0$ and for all $s_0 < s < \tau$.*

Proof. Recall that $v_{jk} = \Psi_{jk}/\beta_{jk}$ form a Riesz basis in $D(K)$ and, therefore, there exist two constants $0 < A \leq B < \infty$ such that

$$A \sum_j \sum_k c^2_{jk} \leq || \sum c_{jk} v_{jk} ||^2 \leq B \sum_j \sum_k c^2_{jk}$$

for all square-summable sequences $\{c_{jk}\}$.

Let $\hat{f}_0$ be the minimax estimator within $D(K)$ and let d^0_{jk} and d_{jk} be the wavelet coefficients of $K\hat{f}_0$ and Kf respectively. Then, from (4)

$$\begin{aligned} R(\hat{f}^{VWD}_\lambda; f) &= E||\hat{f}^{VWD}_\lambda - f||^2 = E|| \sum_j \sum_k \left(\delta_\lambda(\hat{d}_{jk}) - d_{jk}\right) \beta_{jk} v_{jk}||^2 \\ &= E||\sum_j \sum_k \left(\delta_{\beta_{jk}\lambda}(\beta_{jk}\hat{d}_{jk}) - \beta_{jk} d_{jk}\right) v_{jk}||^2 \leq BE \sum_j \sum_k \left(\delta_{\beta_{jk}\lambda}(\beta_{jk}\hat{d}_{jk}) - \beta_{jk} d_{jk}\right)^2 \end{aligned}$$

The $\beta_{jk}\hat{d}_{jk}$'s are independent normal variables with means $\beta_{jk}d_{jk}$ and variances $\sigma^2\beta^2_{jk}$. Then, slightly modifying the results of Donoho & Johnstone (1992, 1995) for the case of non-equal variances, it follows that there exists λ^* such that $\delta_{\beta_{jk}\lambda^*}(\beta_{jk}\hat{d}_{jk})$'s are near-minimax estimators of $\beta_{jk}d_{jk}$'s over the whole range of Besov scale with $s' + 1/2 - 1/p' > 0$ and, therefore, $R_{VWD} = \sup_{f \in B^s_{p,q}\ (C_0)} R(\hat{f}^{VWD}_{\lambda^*}; f) = \sup_{f: Kf \in B^{s'}_{p',q'}\ (C'_0)} R(\hat{f}^{VWD}_{\lambda^*}; f)$ satisfies

$$\begin{aligned} R_{VWD} &\leq B \sup_{d_{jk}} E \sum_j \sum_k \left(\delta_{\beta_{jk}\lambda^*}(\beta_{jk}\hat{d}_{jk}) - \beta_{jk}d_{jk}\right)^2 \leq c_2 \log(n) B \sup_{d_{jk}} E \sum_j \sum_k \left((d^0_{jk} - d_{jk})\beta_{jk}\right)^2 \\ &\leq c_2 \log(n) \frac{B}{A} \sup_{d_{jk}} E|| \sum_j \sum_k (d^0_{jk} - d_{jk}) \beta_{jk} v_{jk}||^2 = c_2 \frac{B}{A} \log(n) R^* = C_2 \log(n) R^* \qquad \Box \end{aligned}$$

The immediate important advantage of VWD is the fact that in contrast to WVD, due to the orthogonality of wavelet basis, the $d_{jk}(y)$'s in (10) are uncorrelated and, hence, thresholding is performed within white noise. Thus, the properly chosen λ^* in VWD implies asymptotic near-optimality of the thresholding procedure (Donoho & Johnstone, 1992, 1995) and, therefore, $c_2 \leq C_1$ in (19). However, since f itself in (9) is expanded in a non-orthogonal vaguelette basis, the Parseval Identity does not hold that yields an additional term B/A in C_2.

6 Conclusions.

Summarizing the main results, we should emphasize several aspects.

Windowed SVD provides a minimax linear estimator over classes of functions of given regularity s. Moreover, for "spatially homogeneous" functions the SVD-estimator is a minimax estimator among all estimators. However, the optimal weights depend on the regularity of the unknown function, which usually cannot be specified in advance.

WVD and VWD yield near-minimax estimators (within the logarithmic factor) but which are near-optimal in the whole range of spaces of different regularities. The logarithmic factor is a price for such "spatial adaptivity". Both estimators outperform SVD and any linear estimator for "spatially inhomogeneous" functions. The main reason lies in the globalized Fourier-like eigenfunctions of the operator K^*K used in SVD for the expansion of the unknown function, while in WVD and VWD the expansion is performed in terms of localized wavelet/vaguelette bases that can parsimoniously reflect the local behavior of the response function.

The examples considered in Section 4 show that the proposed VWD is comparable or even slightly better than WVD. In addition, it is important to note that VWD has definite computational advantages over WVD. The VWD uses the standard DWT that requires $O(n)$ operations and does not need any special software. Performing DVT in WVD demands modifications of the usual software and may be computaionally expensive in the general case. Even for homogeneous operators it requires $O(n \log^2 n)$ operations.

References

[1] Chui, C.K. (1992). *An Introduction to Wavelets*, Academic Press.

[2] Daubechies, I. (1992). *Ten Lectures on Wavelets*, Philadelphia: SIAM.

[3] Donoho, D.L. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harm. Anal.*, to be published.

[4] Donoho, D.L. & Johnstone, I.M. (1992). Minimax estimation via wavelet shrinkage. Technical Report No. 402, Department of Statistics, Stanford University.

[5] Donoho, D.L. & Johnstone, I.M. (1994). Ideal spatial adaption by wavelet shrinkage. *Biometrika* **81**, 425-455.

[6] Donoho, D.L. & Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, to be published.

[7] Donoho, D.L., Johnstone I.M., Kerkyacharian G. & Picard, D. (1995). Wavelet shrinkage: asymptotia? (with discussion). *J. Roy. Statist. Soc., Ser.* **B 57**, 301-369.

[8] Goldenshluger, A. & Nemirovski, A. (1994). On spatial adaptive estimation of nonparametric regression. Technical Report No. 5/94, Faculty of Engineering and Management, Technion - Israel Institute of Technology.

[9] Johnstone, I.M. & Silverman, B.W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18**, 251-280.

[10] Johnstone, I.M. & Silverman, B.W. (1991). Discretization effects in statistical inverse problems. *J. Complexity* **7**, 1-34.

[11] Kolaczyk, E.D. (1994). Wavelet methods for the inverse of certain homogeneous linear operators in the presence of noisy data. Technical Report No. 460, Department of Statistics, Stanford University.

[12] Lepskii O. (1990). On a problem of adaptive estimation in gaussian white noise. *Theory Probab. Applic.* **35**, 454-466.

[13] Mallat, S.G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattn. Anal. Mach. Intell.* **11**, 674-693.

[14] Meyer, Y. (1992). *Wavelets and Operators*, Cambridge University Press.

[15] Nason, G.P. (1993). The **wavethresh** package: wavelet transform and thresholding, available from the StatLib archive.

[16] Nason, G.P. & Silverman, B.W. (1994). The discrete wavelet transform in S. *J. Comp. Graph. Statist.* **3**, 163-191.

[17] O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Stat. Sci.* **1**, 502-527.

[18] Tikhonov, A.N. & Arsenin, V.Y. (1977). *Solutions of Ill-posed Problems*, New York: Wiley.

Table 1. AMSE for WSVD, WVD and VWD estimators

	SNR	WSVD	WVD (D_4)	VWD (D_5)	WVD (D_6)	VWD (D_7)
Bumps	150	0.0162	0.0072	0.0074	0.0075	0.0073
	100	0.0364	0.0147	0.0153	0.0155	0.0149
	75	0.0551	0.0242	0.0251	0.0255	0.0245
	50	0.0888	0.0477	0.0488	0.0500	0.0483
	25	0.1686	0.1377	0.1374	0.1402	0.1360
Blocks	150	0.0260	0.0239	0.0234	0.0241	0.0237
	100	0.0347	0.0381	0.0373	0.0384	0.0372
	75	0.0428	0.0496	0.0488	0.0509	0.0486
	50	0.0522	0.0676	0.0668	0.0735	0.0675
	25	0.0733	0.1086	0.1107	0.1307	0.1046
HeaviSine	150	0.0022	0.0028	0.0027	0.0024	0.0021
	100	0.0027	0.0036	0.0035	0.0034	0.0030
	75	0.0031	0.0043	0.0042	0.0043	0.0038
	50	0.0039	0.0058	0.0056	0.0060	0.0053
	25	0.0054	0.0105	0.0101	0.0100	0.0096