ON DENOISING VIA PENALIZED LEAST-SQUARES RULES

Erik Gudmundson and Petre Stoica*

Uppsala University, Department of Information Technology, Box 337, SE-751 05 Uppsala, Sweden

ABSTRACT

Penalized least-squares (PELS) rules for signal denoising can be obtained via the use of various information criteria (AIC, BIC, etc.) or various minmax LS approaches. Let S denote the set of "significant" parameters in the denoising problem (which is to be determined), let n_S be the dimension of S, and let $n_S\rho$ denote the penalty term of a PELS criterion. We show that, depending on the expression for ρ, the following cases can occur: type-1) If ρ does not depend on S, then denoising via the corresponding PELS rule is equivalent to simple thresholding; and type-2) If ρ depends on n_S only, then the equivalence to thresholding no longer holds but the PELS rule can still be implemented quite efficiently. We also show that the use of BIC leads to an existing PELS rule of type-1 when the noise variance in the denoising problem is known, and to a novel PELS rule of type-2 when the noise variance is unknown.

Index Terms— Signal denoising, thresholding, model order selection, information criterion

1. INTRODUCTORY REMARKS AND PROBLEM FORMULATION

Let $\mathbf{\Phi} \in \mathbb{R}^{N\times N}$ be an orthonormal matrix, and let $\mathbf{\Phi}_S \in \mathbb{R}^{N\times n_S}$ be a matrix constructed from n_S columns of $\mathbf{\Phi}$. Hereafter, $S \subset \{1, ..., N\}$ denotes the set of indices of the columns of $\mathbf{\Phi}$ that form $\mathbf{\Phi}_S$, and n_S is the dimension of S. The assumption that $\mathbf{\Phi}$ is orthonormal is usually made in the literature on the denoising problem, and we adopt it for convenience. Furthermore, we make the common assumption that the data vector, $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$, can be written as:

$$\boldsymbol{y} = \mathbf{\Phi}_S\boldsymbol{\theta}_S + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{\theta}_S \in \mathbb{R}^{n_S\times 1}$ is a vector of unknown parameters, and $\boldsymbol{e} \in \mathbb{R}^{N\times 1}$ is a normally distributed noise vector with mean zero and covariance matrix equal to $\sigma^2\boldsymbol{I}$.

The denoising problem consists of estimating the noise-free "signal" part, $\mathbf{\Phi}_S\boldsymbol{\theta}_S$, of (1) from the noisy data in $\boldsymbol{y}$. Estimation of S is the central part of this problem. Indeed, if S were known then the least-squares (LS) estimate of $\boldsymbol{\theta}_S$, namely

$$\hat{\boldsymbol{\theta}}_S = \mathbf{\Phi}_S^T\boldsymbol{y} \tag{2}$$

($(\cdot)^T$ denotes the transpose) could simply be used to produce the following "signal" estimate: $\hat{\boldsymbol{y}} = \mathbf{\Phi}_S\hat{\boldsymbol{\theta}}_S$.

Methods for estimating S have been suggested by several authors, see e.g. [1–5] and the references therein. Estimating S is also the main topic of this paper. As in most of the cited references, we assume that $n_S \ll N$. While this requirement is a limitation, we believe it is a fairly natural one: if the value of n_S was comparable with N, then most denoising methods, if not all, would likely perform rather poorly, and therefore this situation appears to be of relatively minor practical interest.

Under the assumption that σ^2 in (1) is known, we use the Bayesian Information Criterion (BIC), see e.g. [6, 7] and references in [7], to obtain a denoising rule of the penalized least-squares (PELS) type, which we call DEBIT (DEnoising via BIc Thresholding). We show that DEBIT is equivalent to simple thresholding, and consequently we compare it with a thresholding rule in [1] (obtained by a different approach). DEBIT has also been obtained in [2], in a related context, by means of a minimum description length (MDL) approach, and in a more general form in [4] by using an extended BIC approach.

When σ^2 in (1) is unknown, the use of BIC is shown to lead to a slightly more involved PELS rule. While this rule is no longer equivalent to thresholding, its use is still quite simple computationally. Using a slight modification of the BIC, we derive the DEMBIT (DEnoising via Modified BIc Thresholding) method, that appears to be a useful addition to the set of existing denoising rules, since it does not require that σ^2 be known and since it is relatively simple to implement (as already mentioned).

We will also discuss briefly the KIC$_c$ and MDL methods derived in [5], which we will use in a comparison study.

Finally, we will compare numerically the denoising performance of some of the thresholding rules in [1] and [5], of DEBIT, and of DEMBIT.

2. PELS DENOISING RULES

Under the assumptions stated earlier, the likelihood function of $\boldsymbol{y}$ in (1) is given by:

$$f_S(\boldsymbol{y}|\boldsymbol{\theta}_S, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\left\|\boldsymbol{y} - \mathbf{\Phi}_S\boldsymbol{\theta}_S\right\|^2\right\}. \tag{3}$$

As is well known, the maximum likelihood estimates of $\boldsymbol{\theta}_S$ an σ^2 (which maximize (3), for a given S) coincide with the LS estimates of these parameters: (2) for $\hat{\boldsymbol{\theta}}_S$, and

$$\hat{\sigma}^2 = \frac{1}{N}\left\|\boldsymbol{y} - \mathbf{\Phi}_S\hat{\boldsymbol{\theta}}_S\right\|^2 \tag{4}$$

for σ^2. In the sequel, we first assume that σ^2 is known, and then we go on to relax this assumption.

2.1. DEBIT

Assuming that σ^2 is *known*, the BIC estimate of S is given by ([6,7] etc.):

$$\min_S[-2\ln f_S(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_S, \sigma^2) + n_S \ln N]. \tag{5}$$

Some simple calculations show that (5) can be reduced to the minimization of the following PELS criterion:

$$\min_S\left[\left\|\boldsymbol{y} - \mathbf{\Phi}_S\hat{\boldsymbol{\theta}}_S\right\|^2 + n_S\rho\right]; \qquad \rho = \sigma^2\ln N. \tag{6}$$

This work was partly supported by the Swedish Science Council (VR).
* Corresponding author. E-mail: erik.gudmundson@it.uu.se.

Next we note that

$$\left\|\boldsymbol{y}-\boldsymbol{\Phi}_S\hat{\boldsymbol{\theta}}_S\right\|^2=\left\|(\boldsymbol{I}-\boldsymbol{\Phi}_S\boldsymbol{\Phi}_S^T)\boldsymbol{y}\right\|^2=\|\boldsymbol{y}\|^2-\left\|\hat{\boldsymbol{\theta}}_S\right\|^2 \quad (7)$$

which implies that (6) is equivalent to:

$$\max_S \sum_{k=1}^{n_S}(|\hat{\theta}_S^k|^2-\rho) \quad (8)$$

(hereafter, $\hat{\theta}_S^k$ denotes the k-th component of $\hat{\boldsymbol{\theta}}_S$). The solution to the maximization problem above is given by:

$$\hat{S}=\left\{k\in\{1,...,N\}\,|\,|\hat{\theta}^k|>\sqrt{\rho}\right\}; \qquad \hat{\boldsymbol{\theta}}=\boldsymbol{\Phi}^T\boldsymbol{y}, \quad (9)$$

where $\hat{\theta}^k$ is the k-th element in $\hat{\boldsymbol{\theta}}$. To see this, we observe that for $S=\hat{S}$ all the terms in (8) are positive; and that if we remove elements from $\hat{S}$ in (9) or add elements to it, then the corresponding value of the criterion in (8) becomes smaller than its value at $S=\hat{S}$ due to omitting positive terms from it or, respectively, including negative terms in it.

It follows from the previous analysis that the application of BIC to the model in (1) leads to the simple thresholding rule in (9) with a threshold $\sqrt{\rho}=\sigma\sqrt{\ln N}$. This denoising rule, which we call DEBIT, was previously derived in [2] by means of an MDL approach and of a limiting argument; it was also obtained in [4], for a slightly more general problem, by using an extended BIC approach. Interestingly, DEBIT is similar to the thresholding-based denoising rule introduced in [1], using a deterministic minmax LS approach, for which $\hat{S}$ is also given by (9) but with $\sqrt{\rho}=\sigma\sqrt{2\ln N}$. We will refer to the latter rule as THRESH.

To conclude this sub-section, we remark on the fact that the derivation of (9) remains valid for any PELS rule of the form of (6), for which ρ is independent of S. For instance, the use of AIC, in lieu of BIC, would lead to (6) with $\rho=2\sigma^2$. Because BIC has more appealing properties than AIC and than other IC rules (at least, for sufficiently large values of N; see e.g. [7] and the references therein), we will not consider any of these possible alternative forms of (6) in this paper.

2.2. DEMBIT

When σ^2 is *unknown*, the BIC rule is still given (under the previously made assumptions) by (5) but with σ^2 replaced by $\hat{\sigma}^2$ in (4). As indicated above, the threshold used in THRESH is larger than that of DEBIT by a factor of $\sqrt{2}$; therefore THRESH can be seen as method that penalizes large values of n_S harder than BIC. Because simulations have shown that THRESH performs better than DEBIT, we modify (5) so that it uses the THRESH penalty:

$$\min_S[-2\ln f_S(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_S,\hat{\sigma}^2)+n_S 2\ln N]. \quad (10)$$

A simple calculation shows that the so-modified rule estimates S as the solution to the following minimization problem:

$$\min_S\left[\ln\left\|\boldsymbol{y}-\boldsymbol{\Phi}_S\hat{\boldsymbol{\theta}}_S\right\|^2+2\gamma_S\right]; \qquad \gamma_S=\frac{n_S\ln N}{N}. \quad (11)$$

Asymptotically in N, using the fact that for $x\ll 1$

$$\ln(1+x)\approx x, \quad (12)$$

the PELS rule above is equivalent to

$$\min_S\left[\ln\left\|\boldsymbol{y}-\boldsymbol{\Phi}_S\hat{\boldsymbol{\theta}}_S\right\|^2+\ln(1+2\gamma_S)\right]. \quad (13)$$

Making use of (7) we can re-write (13) as

$$\max_S\left[(1+2\gamma_S)\|\hat{\boldsymbol{\theta}}_S\|^2-2\gamma_S\|\boldsymbol{y}\|^2\right]. \quad (14)$$

In general, the solution to the above maximization problem can no longer be obtained via thresholding. Nevertheless, it can still be efficiently computed as follows:

- Let $\hat{\boldsymbol{\theta}}=\boldsymbol{\Phi}^T\boldsymbol{y}$ (as in (9)), and observe that $\hat{\boldsymbol{\theta}}_S$ is made from the elements of $\hat{\boldsymbol{\theta}}$ whose indexes belong to the set S. Order the entries of $\hat{\boldsymbol{\theta}}$ in a decreasing magnitude order:
$$\left|\hat{\theta}_{\nu_1}\right|\geqslant\left|\hat{\theta}_{\nu_2}\right|\geqslant\cdots\geqslant\left|\hat{\theta}_{\nu_N}\right|,$$
where $\nu_k\in\{1,\ldots,N\}$, and $\nu_j\neq\nu_k$ if $j\neq k$.
- For each fixed value of n_S ($n_S=1,2,\ldots$), the corresponding optimum solution of (14) is clearly given by:
$$\tilde{S}_{n_S}=\{\nu_1,\ldots,\nu_{n_S}\}. \quad (15)$$
- To choose one of the sets above (see (15)), we could compare the values of the approximate criterion in (14) corresponding to $S=\tilde{S}_{n_S}$, for $n_S=1,2,\ldots,N$. However, to avoid the errors introduced by the approximation in (12), we use instead the exact criterion (11): we evaluate (11) with $S=\tilde{S}_{n_S}$, for $n_S=1,2,\ldots,N$, and choose the set that gives the smallest value.

The so-obtained denoising rule will be called DEMBIT. Evidently, the DEMBIT algorithm outlined above can be used to find the solution of any PELS problem that has the form of (13) with a general "penalty" factor γ_S, provided that γ_S depends only on n_S.

The DEMBIT denoising rule, introduced above, appears to be novel. Its main appeal lies in the fact that it does not require σ^2 to be known, unlike most other existing denoising rules, and that its implementation is relatively simple computationally.

2.3. MDL and KIC$_c$ denoising

Closely related to the DEMBIT denoising rule are the minimum description length (MDL) [2] and the Kullback Information Criterion corrected (KIC$_c$) [5] denoising methods. The principles of these rules are similar to the one of DEMBIT, but these rules differ in the derivation and in the minimizing criteria. For MDL denoising, let k_{opt} denote the $k=1,...,N$ that minimizes

$$\mathrm{MDL}(k)=(N-k)\ln\left(\frac{\|\boldsymbol{y}\|^2-\|\boldsymbol{y}_k\|^2}{N-k}\right)+ k\ln\left(\frac{\|\boldsymbol{y}_k\|^2}{k}\right)-\ln\left(\frac{k}{N-k}\right), \quad (16)$$

where $\boldsymbol{y}_k$ is the vector consisting of the k elements of $\boldsymbol{y}$ that have the largest magnitude. For KIC$_c$ the criterion that gives k_{opt} is defined as

$$\mathrm{KIC}_c(k)=N\ln\left(\frac{\|\boldsymbol{y}\|^2-\|\boldsymbol{y}_k\|^2}{N}\right)+2\frac{(k+1)N}{N-k-2}- N\psi\left(\frac{N-k}{2}\right), \quad (17)$$

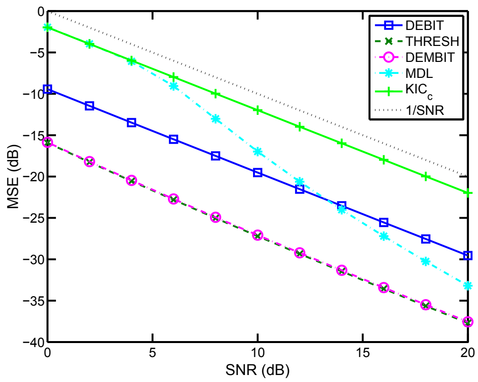


Fig. 1. MSE as a function of the SNR when $n_{S_0} = 5$.

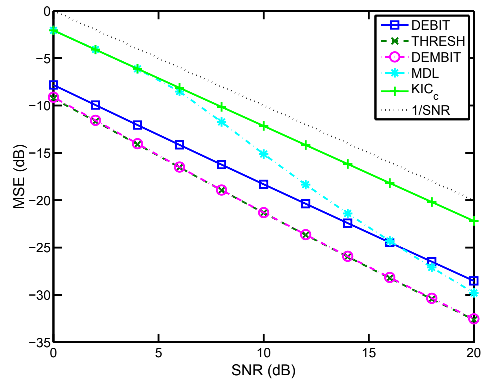


Fig. 2. MSE as a function of the SNR when $n_{S_0} = 20$.

where $\psi(\cdot)$ is the digamma function [8]. Now, let

$$\boldsymbol{y}_{k_{opt}} = \begin{bmatrix} y_{\nu_1} & \dots & y_{\nu_{k_{opt}}} \end{bmatrix}^T .$$

Then $\hat{S} = \{\nu_1, ..., \nu_{k_{opt}}\}$. For a detailed description of these methods, see [5].

3. NUMERICAL EXAMPLES AND CONCLUDING REMARKS

We will consider two cases. In the first one we randomly generate a sparse vector $\boldsymbol{\theta}$ and use an identity matrix, $\boldsymbol{I}$, as the regressor matrix $\boldsymbol{\Phi}$. In the second case we use a Daubechies type mother wavelet as regressor matrix for three different signals: a chirp, a seismic signal, and a piecewise polynomial signal with discontinuity. These examples were previously studied in [5]. To generate the wavelets we use the WaveLab framework [9].

3.1. Identity Regressor Matrix

We consider a case in which $\boldsymbol{\Phi} = \boldsymbol{I}$, such that $\boldsymbol{\Phi}_S \boldsymbol{\theta}_S$ in (1) can be viewed as the "signal" itself (rather than a linear model thereof). We choose $N = 500$, and consider two values for n_{S_0}: $n_{S_0} = 5$ and $n_{S_0} = 20$; here, S_0 denotes the true set used to produce the data. We generate S_0 randomly by picking up (without replacement) n_{S_0} indices from a uniform distribution over the set $\{1, \dots, N\}$. The elements of $\boldsymbol{\theta}_{S_0}$ are independently drawn from n_{S_0} zero-mean normal distributions with standard deviation $\sigma_\theta = 1$. Finally, the noise variance σ^2 is selected to make the signal-to-noise ratio,

$$\text{SNR} = \frac{\|\boldsymbol{\theta}_{S_0}\|^2}{N\sigma^2} \tag{18}$$

vary from 0 dB to 20 dB.

Figures 1 and 2 show the mean squared error (MSE) of the methods outlined in the previous section,

$$E\left[\frac{\left\|\boldsymbol{\Phi}_{S_0}\boldsymbol{\theta}_{S_0} - \boldsymbol{\Phi}_{\hat{S}}\hat{\boldsymbol{\theta}}_{\hat{S}}\right\|^2}{\|\boldsymbol{\theta}_{S_0}\|^2}\right] \tag{19}$$

for $n_{S_0} = 5$ and, respectively, $n_{S_0} = 20$. The expectation in (19) is empirically evaluated using 10^6 Monte-Carlo simulation runs corresponding to (100 realizations of S_0) $\times$ (100 realizations of $\boldsymbol{\theta}_{S_0}$) $\times$ (100 realizations of $\boldsymbol{e}$). As a comparison, the naive LS estimate of $\boldsymbol{\theta}$, whose MSE is equivalent to $1/\text{SNR}$, is also plotted.

We can see that the proposed method DEMBIT and the THRESH method perform similarly and outperform the other methods. Bear in mind, though, that THRESH is supplied with the true noise variance and thus assumes it to be known, a rather unlikely assumption in real applications. KIC$_c$ and MDL, like DEMBIT, estimate the noise variance from the data but they perform worse.

3.2. Wavelets as Regressor Matrix

In this subsection a Daubechies type wavelet of length 6 and lower resolution cutoff 3 is used as regressor matrix. Let $\boldsymbol{f} = [f(t_1) \ \dots \ f(t_N)]^T$ be the vector containing the time samples of the noise-free signal. As in [5], the following three different signals are used:

1) A chirp signal,

$$f_{chirp}(t) = \sin\left(40\pi(1.5t^2 - 1.36t + 0.68)\right), \quad t \in [0, 1]$$

to which noise with variance $\sigma_0^2 = 0.5$ is added. This type of signal is typically encountered in radar and sonar signal processing. The wavelet coefficients (i.e. the vector $\boldsymbol{\theta}$) of the noise-free signal can be seen in Figure 3(a).

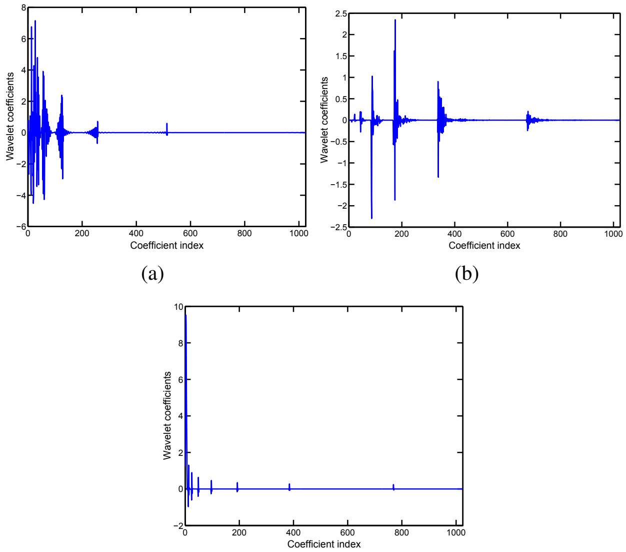
2) A 1-D seismic signal, $f_{seis}(t)$, with added noise variance of $\sigma_0^2 = \alpha \|\boldsymbol{f}\|_\infty$, $\alpha = 0.01$. (i.e., the noise is proportional to the magnitude of the largest element of the noise-free seismic data). This example signal is included in the WaveLab package and is used as a test data set in the seismic industry [5]. A plot of the wavelet coefficients of the noise-free signal is shown in Figure 3(b).

3) A piecewise polynomial signal,

$$f_{PP}(t) = \begin{cases} 4t^2(3-4t) & t \in [0.00, 0.50] \\ \frac{4}{3}t(4t^2 - 10t + 7) - \frac{3}{2} & t \in]0.50, 0.75] \\ \frac{16}{3}t(t-1)^2 & t \in]0.75, 1.00] \end{cases}$$

to which noise with variance $\sigma_0^2 = 0.1$ is added. This is an example of a smooth signal with a discontinuity, often encountered in edge detection applications. The wavelet coefficients of the noise-free signal can be seen in Figure 3(c).

All signals consisted of 1024 time samples corrupted with white Gaussian noise of variance σ_0^2. To transform the signal to the wavelet domain we use the package WaveLab850 [9].



(a) (b) (c)

Fig. 3. The wavelet coefficients of (a) the chirp signal; (b) the seismic signal; and (c) the piecewise polynomial signal.

The MDL, KIC$_c$, and DEMBIT methods are again compared using the MSE of the denoised signal $\hat{\boldsymbol{f}}$

$$E\left[\left\|\boldsymbol{f}-\hat{\boldsymbol{f}}\right\|^2\right], \tag{20}$$

where $\boldsymbol{f}$ is the noise-free signal. The expectation is empirically evaluated using Monte-Carlo simulations, but this time with 1000 different noise realizations. The results are shown in Table 1. We see that the performance depends on the signal type. For the piecewise polynomial signal, the proposed method DEMBIT outperforms both MDL and KIC$_c$, whereas for the other signal types KIC$_c$ outperforms the other methods. As can be seen in Figure 3(c), the piecewise polynomial signal has a very sparse structure and this appears to be the main reason why DEMBIT performs best. The sparseness for the seismic signal (Figure 3(b)) and the chirp signal (Figure 3(a)) is smaller and KIC$_c$, which penalizes less than DEMBIT and MDL, shows the best performance. Note, though, that the differences in performance for these signals are small (on the order of 1 dB) whereas for the piecewise polynomial signal, DEMBIT outperforms KIC$_c$ by about 3.5 dB.

4. CONCLUDING REMARKS

In this paper we have derived a new approach for the denoising of sparse signals using a penalized least squares criterion. Methods have been derived for both known and unknown noise variances, giving rise to simple yet powerful denoising methods. We have studied the performance of the methods using Monte-Carlo simulations and compared them to previously known methods. Our conclusion is that the proposed method, DEMBIT, performs best for cases where the data structure is very sparse, and that its performance is not much lower than that of other well-known methods when the data structure has less sparsity.

Table 1. MSE (in dB) of the different denoised signals

Signal	Denoising method		
	DEMBIT	MDL	KIC$_c$
Chirp	-8.49	-7.75	-9.64
Seismic	-24.31	-24.75	-25.22
Piecewise polynomial	-22.36	-18.51	-18.94

5. ACKNOWLEDGEMENT

The authors are grateful to Dr. Maïza Bekara for sharing his code for computing the MDL and KIC$_c$ algorithms, and to Dr. Yngve Selén and Dr. Niclas Sandgren for their comments on this work. Some of the computations were performed using UPPMAX resources under Project p2007002.

6. REFERENCES

[1] D. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, pp. 425–455, 1994.

[2] J. Rissanen, "MDL Denoising," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2537–2543, Nov 2000.

[3] S. Beheshti and M. A. Dahleh, "A new information-theoretic approach to signal denoising and best basis selection," *IEEE Transactions of Signal Processing*, vol. 53, no. 10, pp. 3613–3624, October 2005.

[4] P. Stoica and N. Sandgren, "Total-variance reduction via thresholding: application to cepstral analysis," *IEEE Transactions on Signal Processing*, vol. 55, no. 1, pp. 66–72, January 2007.

[5] M. Bekara, L. Knockaert, A.-K. Seghouane, and G. Fleury, "A model selection approach to signal denoising using Kullback's symmetric divergence," *Signal Processing*, vol. 86, no. 7, pp. 1400–1409, July 2006.

[6] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[7] P. Stoica and Y. Selén, "Model-order selection — A review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, July 2004.

[8] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, 9th dover printing, 10th gpo printing edition, 1964.

[9] "WaveLab850," [Online.] Available: `http://www-stat.stanford.edu/~wavelab/`, Sept. 2007.