

2-1-2009

Wavelet Reconstruction of Nonuniformly Sampled Signals

Leming Qu
Boise State University

Partha S. Routh
Seismic Technology Development, ConocoPhillips, Houston, TX

Phil D. Anno
Seismic Technology Development, ConocoPhillips, Houston, TX

Wavelet Reconstruction of Nonuniformly Sampled Signals

Leming Qu, Partha S. Routh, and Phil D. Anno

Abstract—For the reconstruction of a nonuniformly sampled signal based on its noisy observations, we propose a level dependent l_1 penalized wavelet reconstruction method. The LARS/Lasso algorithm is applied to solve the Lasso problem. The data adaptive choice of the regularization parameters is based on the AIC and the degrees of freedom is estimated by the number of nonzero elements in the Lasso solution. Simulation results conducted on some commonly used 1-D test signals illustrate that the proposed method possesses good empirical properties.

Index Terms—AIC, LARS, Lasso, wavelet.

I. INTRODUCTION

MODERN digital data processing of signals always uses a discretized and most likely the noisy version of the original signal f that is obtained by sampling f on a discrete set. The question then arises is how f can be reconstructed from its samples. The problem can be stated as: given noisy data at sampling points $\{t_1, \dots, t_n\}$

$$y_i = f(t_i) + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d. } \sim N(0, \sigma^2)$$

we wish to estimate the unknown signal f from the noisy sample. Without loss of generality, assume that the function f is defined on $[0, 1]$. When t_1, \dots, t_n is uniform (a regular grid), the standard methods based on Fourier analysis or wavelet analysis are well studied. We are concerned with the nonuniform samples in this letter. The problem of signal reconstruction from nonuniformly sampled data arises in many applications, including sampling systems with sampling jitter, the reconstruction of signals from missing samples, and the processing of geophysical data.

The setting of the problem is the same as the one in [2] and [3] by assuming further that we have many missing samples on a fine regular grid. That is, $t_i = n_i/m$, for some n_i and some fine grid number $m = 2^J$ with resolution J that is determined by users. Usually, $m \geq n$ so that the approximation errors by moving nondyadic points to dyadic points are negligible. Let \mathbf{f}_m be the underlying signal collected at all dyadic points $\{i/m, i = 1, \dots, m\}$. Let \mathbf{W} be a given orthogonal discrete wavelet transform (DWT) and $\boldsymbol{\theta} = \mathbf{W}\mathbf{f}_m$ be the wavelet

transform of \mathbf{f}_m , then $\mathbf{f}_m = \mathbf{W}^T\boldsymbol{\theta}$, where \mathbf{W}^T is the inverse discrete wavelet transform (IDWT).

By modeling the unknown signal \mathbf{f}_m with its DWT $\boldsymbol{\theta}$, we overparameterized the linear model derived from the observations $\mathbf{y} = [y_1, y_2, \dots, y_n]^T$. For a function f in wide classes such as Besov space, its wavelet coefficients are sparse. This prior knowledge is used to regularize the overparameterized linear model.

Let \mathbf{S} be the $n \times m$ sampling matrix which extracts the $\{n_1, \dots, n_m\}$ th entry from \mathbf{f}_m , then $\mathbf{f} = \mathbf{S}\mathbf{f}_m$ is the discrete signal at the sampling points. Denote $\mathbf{A} = \mathbf{S}\mathbf{W}^T$, then $\mathbf{f} = \mathbf{A}\boldsymbol{\theta}$. The observed data can be expressed as a linear model

$$\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$$

where $\boldsymbol{\varepsilon}$ is the noise vector. Antoniadis and Fan [2] studied several aspects of the penalized least-squares problem

$$\min_{\boldsymbol{\theta}} 2^{-1}\|\mathbf{y} - \mathbf{A}\boldsymbol{\theta}\|^2 + \lambda \sum_{i=1}^m p(|\theta_i|)$$

for a given penalty function $p(\cdot)$ and regularization parameter $\lambda > 0$. They proposed a two-step procedure to solve the above minimization problem for a wide choices of penalty function $p(\cdot)$. The first step is to find the nonlinear regularized Sobolev interpolators (NRSI), and the second step is to find the regularized one-step estimators (ROSE). The risk behavior of ROSE is quantified. They also introduced the graduated nonconvexity algorithm to handle penalized least-squares problems when the penalty function $p(\cdot)$ is nonconvex. One drawback of the two-step procedure for ROSE is that in the first step, the NRSI needs the smoothing Sobolev interpolation parameter s which choice is not easily determined automatically. For nonconvex penalty, it is in general very costly to find an (approximate) solution in this case.

In [1], the penalty function $p(\cdot)$ is taken as the sum of block-wise reproducing kernel Hilbert space norms. Their simulation results show superior performance in comparison with some existing methods.

In this letter, we choose the penalty function $p(\cdot)$ as the level dependent l_1 norm. That is, we solve the following particular penalized least-squares problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}} 2^{-1}\|\mathbf{y} - \mathbf{A}\boldsymbol{\theta}\|^2 + \tau \left(\sum_{k=1}^{2^L} \alpha |\theta_{L-1,k}| + \sum_{j=L}^{J-1} \sum_{k=1}^{2^j} 2^{js} |\theta_{j,k}| \right) \\ = \min_{\boldsymbol{\theta}} 2^{-1}\|\mathbf{y} - \mathbf{A}\boldsymbol{\theta}\|^2 + \tau \|\mathbf{D}\boldsymbol{\theta}\|_1 \quad (1) \end{aligned}$$

for a given $s \geq 0$ and $\tau > 0$, where $\theta_{L-1,k}$ is the discrete scaling coefficient at the coarsest level L and location k , $\theta_{j,k}$ is the discrete wavelet coefficient at level j and location k . Denote $\boldsymbol{\theta}_{L-1}$ the 2^L -vector of scaling coefficients and $\boldsymbol{\theta}_j$ the 2^j -vector

Manuscript received June 18, 2008; revised October 11, 2008. Current version published January 08, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Erik G. Larsson.

L. Qu is with the Department of Mathematics, Boise State University, Boise, ID 83725 USA (e-mail: lqu@boisestate.edu).

P. S. Routh and P. D. Anno are with Seismic Technology Development, ConocoPhillips, Houston, TX 77252-2197 USA (e-mail: Partha.S.Routh@ConocoPhillips.Com; phil.d.anno@conocophillips.com).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>

Digital Object Identifier 10.1109/LSP.2008.2008939

of discrete wavelet coefficients at the resolution level j , then $\boldsymbol{\theta} = [\boldsymbol{\theta}_{L-1}^T, \boldsymbol{\theta}_L^T, \dots, \boldsymbol{\theta}_{J-1}^T]^T$ is the DWT coefficients of \mathbf{f} . \mathbf{D} is a $m \times m$ diagonal matrix denoting the weights of $\boldsymbol{\theta}$. Its diagonal elements are α for $\boldsymbol{\theta}_{L-1}$ and 2^{js} for $\boldsymbol{\theta}_j$. The weight for the scaling coefficients at the lowest resolution level, α , is set at a small value (in orthogonal design case $\mathbf{A} = \mathbf{W}^T$, the weight for the scaling coefficients at the coarsest level is usually zero).

Level dependent thresholding for equally spaced data is well-known in wavelet denoising [4], yet it has not been used in the unequally spaced data denoising problem. The term 2^{js} in the penalty function imposes the prior belief that the discrete wavelet coefficients of the signals to be reconstructed decay exponentially fast with increasing resolution level j , which is the characteristic of functions in Besov space (see Section II).

We discuss the application of some existing fast algorithm to solve (1). The important issue of the data adaptive choice of the regularization parameter $\lambda \equiv (s, \tau)$ is addressed. Note that our method requires no specification of unknown problem-specific parameters, neither s nor τ . This differs from the two-step procedure in [2] which requires a manual specification of s in the first step.

II. WAVELET TRANSFORM

In contrast to Fourier representations, wavelets provide economical representations for a wide class of signals including signals with discontinuities.

The 1-D wavelet transform represents a 1-D continuous-time signal $f(t)$, in terms of shifted versions of a scaling function ϕ and shifted and dilated versions of a mother wavelet function ψ [10]. For special choices of ϕ and ψ , the functions at resolution level j and time position $k/2^j$

$$\phi_{j,k}(t) = 2^{j/2} \phi(2^j t - k), \quad j, k \in Z$$

and

$$\psi_{j,k}(t) = 2^{j/2} \psi(2^j t - k), \quad j, k \in Z$$

form an orthonormal basis. Furthermore, the periodized version of this basis (denoted by $\tilde{\phi}$ and $\tilde{\psi}$ in the following) define a multiresolution analysis of $L^2([0, 1])$. For any $f(t) \in L^2([0, 1])$, the following standard orthonormal wavelet expansion holds:

$$f(t) = \sum_{k=1}^{2^{j_0}} \alpha_{j_0,k} \tilde{\phi}_{j_0,k}(t) + \sum_{j=j_0}^{\infty} \sum_{k=1}^{2^j} \beta_{j,k} \tilde{\psi}_{j,k}(t)$$

with the scaling coefficients $\alpha_{j,k} = \int_0^1 f(t) \tilde{\phi}_{j,k}(t) dt$ and wavelet coefficients $\beta_{j,k} = \int_0^1 f(t) \tilde{\psi}_{j,k}(t) dt$.

For discrete-time signal \mathbf{f}_m with an equally spaced $m = 2^J$ samples, the m wavelet coefficients can be efficiently computed by cascade algorithm in $O(m)$ operations via a filterbank consisting of lowpass filters, highpass filters, upsamplers, and decimators [10]. By periodization of the input discrete signal \mathbf{f}_m , the Discrete wavelet transform can be expressed in matrix form as $\boldsymbol{\theta} = \mathbf{W} \mathbf{f}_m$ where \mathbf{W} is an orthonormal matrix with entries either zeros or nonzeros related to the filter coefficients [14, p. 115].

Wavelets provide economical representations for signals in wide classes such as Besov spaces. Roughly speaking, a Besov space $B_{p,q}^s$ contains functions with “ s derivatives in L_p ” with q measuring finer smoothness distinctions. Hence, s indicates the

degree of smoothness of the underlying signal f . Besov spaces with different s , p , and q characterize many classes of signals in addition to L_2 -Sobolev space signals; for example, in 1-D, $B_{1,1}^1$ contains piece-wise polynomial signals [10]. Assume that the signal f is in a Besov ball. Because of simple characterization of this space via the wavelet coefficients of its members for sufficiently smooth wavelet basis functions, the unit ball of a Besov space $B_{p,q}^s([0, 1])$ can be defined as

$$B_{p,q}^s = \left\{ f \in L_p : \sum_j \left(2^{j(s+1/2-1/p)} \|\boldsymbol{\beta}_j\|_p \right)^q \leq 1 \right\}$$

where $\boldsymbol{\beta}_j$ is the 2^j -vector of wavelet coefficients at the resolution level j . Note that the wavelet coefficients in the definition of the Besov space are continuous-wavelet coefficients. They are approximately a \sqrt{m} fraction of the discrete-wavelet coefficients $\boldsymbol{\theta} = \mathbf{W} \mathbf{f}_m$, i.e., $\beta_{jk} = \theta_{jk} / \sqrt{m}$ (see [4, p. 1211]). Hence, the wavelet coefficients of Besov space signals decay exponentially fast with increasing resolution level.

III. LARS FOR LASSO PROBLEM

By a simple rescaling, the problem (1) is the same as the Lasso [11] problem. That is, let $\tilde{\boldsymbol{\theta}} = \mathbf{D} \boldsymbol{\theta}$ and $\tilde{\mathbf{A}} = \mathbf{A} \mathbf{D}^{-1}$, the problem (1) is then in standard Lasso form

$$\min_{\tilde{\boldsymbol{\theta}}} 2^{-1} \|\mathbf{y} - \tilde{\mathbf{A}} \tilde{\boldsymbol{\theta}}\|^2 + \tau \|\tilde{\boldsymbol{\theta}}\|_1. \quad (2)$$

Problems of the Lasso type have become familiar over the past three decades, particularly in statistical and signal processing contexts. See [7] for a review of various algorithms for solving the Lasso problem. We choose a Homotopy algorithm to solve problem (2) because we do not know which $\lambda \equiv (s, \tau)$ to use initially, and we would prefer an algorithm which finds solutions for essentially all nonnegative values of the regularization parameter τ for given s in a single run. The least angle regression (LARS) [6], a new model selection algorithm, is a Homotopy algorithm. The Lasso implementation of the LARS calculates all possible Lasso solution in a single run. For given s , the LARS/Lasso algorithm begins with $\tau = \|\tilde{\mathbf{A}}^T \mathbf{y}\|_{\infty}$, and gradually reduce τ in stages that predictably change the sparsity pattern in $\tilde{\boldsymbol{\theta}}$. Its remarkable efficiency follows from its ability to systematically update the resulting sequence of solutions (see [5] for discussions). This sparsity pattern predictability comes from the fact that $\|\tilde{\boldsymbol{\theta}}\|_1$, when restricted to the subset of nonzero indices of $\tilde{\boldsymbol{\theta}}$, is locally a linear function of $\tilde{\boldsymbol{\theta}}$. This allows one to solve the local problems (for a limited range of τ) analytically and to piece together local solutions to get solutions for all regions of τ . Hence, the solution path for each element of $\tilde{\boldsymbol{\theta}}$ is piecewise linear on τ . At each iteration, a least-squares subproblem that involves a subset of the columns of $\tilde{\mathbf{A}}$ is solved by using the results from the previous iteration. The resulting algorithm generates solutions for all τ with a computational cost of solving a single least-squares problem in the final set of variables. For $n \ll m$ and $\tilde{\boldsymbol{\theta}}$ sparse, Donoho and Tsai [5] argue that LARS/Lasso algorithm is the most efficient way to solve any of the Lasso problems (2).

IV. DATA ADAPTIVE CHOICE OF λ

The regularization parameter s is roughly the number of derivatives of the signal f . The larger the s , the smoother the

f . It is reasonable to set the s in a relative small range. In our simulation, we choose $s \in [0, 2]$.

The regularization parameter τ controls the sparsity of θ . If $\tau > \|\hat{\mathbf{A}}^T \mathbf{y}\|_\infty$, the unique solution of (2) is the zero vector. When τ decreases, the sparsity of the solution decreases, too. The LARS/Lasso algorithm gives a full path of the solution for essentially all τ in a single run.

The task of choosing the right regularization parameter $\lambda \equiv (s, \tau)$ is a difficult and important one. Main approaches in the literature are 1) cross-validation (CV) such as leave-one-out CV or v -fold CV; 2) information criteria such as AIC, BIC; and 3) L-curve. While these approaches in the classical Tikhonov regularization (or ridge regularization in statistics), i.e., the l_2 norm data misfit term and the l_2 norm penalty term, are extensively studied, they are less researched in the Lasso problem, especially in the case when there are many more unknowns than the number of observations ($m > n$).

Fu [8] discussed the CV choice of λ in Lasso problem when $m \leq n$. Zou *et al.* [13] discussed the degrees of freedom (df) in Lasso problem and its use in AIC and BIC. Hennenfent *et al.* [9] discussed L-curve for Lasso problem. We use the information criteria to select the λ . The df estimate of the solution is needed in order to use AIC or BIC.

In Lasso problems with $\hat{\mathbf{A}}$ full column rank (thus, $m \leq n$ is implied), it was proved that the number of nonzero elements in the solution is an unbiased estimator of the df [13]. In $m > n$ case, there is no available proved close form formula for unbiased df estimator. We extend the results from [13] and use the number of nonzero elements in the solution θ_λ as the df estimate, \hat{df}_λ .

Another approach to estimate the df is to use perturbed dataset [12], which is computationally more intensive. We tried this approach in our problem and found it performed similarly to the simple df estimate (the number of nonzero elements in the solution).

Once a good df estimate is obtained, the AIC criteria is

$$\text{AIC}(\lambda) = \frac{\|\mathbf{y} - \hat{\mathbf{f}}_\lambda\|^2}{n\sigma^2} + \frac{2}{n} \hat{df}_\lambda$$

the BIC criteria is to replace $2/n$ in the above equation by $\log(n)/n$. The optimal λ is chosen as the minimizer of $\text{AIC}(\lambda)$ for AIC criteria or of $\text{BIC}(\lambda)$ for BIC criteria.

AIC and BIC possess different asymptotic optimality. It is well-known that AIC tends to select the model with the optimal prediction performance, while BIC tends to identify the true sparse model if the true model is in the candidate list; see [15] and references therein. AIC is more appropriate for our problem, since the sparsity of the model is not our primary concern. Our goal is to reconstruct $\mathbf{f} = \mathbf{A}\theta$, where the θ is merely an intermediate tool to facilitate the reconstruction. We target the model with optimal prediction performance.

In many practical situations, σ is unknown. A reasonably good estimate $\hat{\sigma}$ of σ can be used. The statistics literature includes many proposals. We use the median absolute deviation (MAD) estimate, $\text{median}\{|y_{2i} - y_{2i-1}| / (.6745\sqrt{2}) : i = 1, \dots, n/2\}$, suggested by Donoho and Johnstone in [4].

V. SIMULATION STUDY

This section reports the empirical performance of a simulation study of our level dependent l_1 regularized wavelet

estimator, compared with the wavelet kernel (WK) penalized estimation method proposed in [1]. The MATLAB code carrying out this simulation is available at <http://math.boisestate.edu/~qu>.

The wavelets used in all our simulations are Daubechies extremal phase wavelets with five vanishing moments except that two vanishing moments is used for Blocks signal. The lowest resolution level is set at $L = 1$. We use the WaveLab (<http://www-stat.stanford.edu/~wavelab/>) for DWT and IDWT.

The SolveLasso routine available in SparseLab (<http://sparselab.stanford.edu/>) implements the LARS/Lasso algorithm. We use the SolveLasso to solve Lasso problem (2).

The grid for s is ten equally spaced values in $[0, 2]$. The weight for the scaling coefficients α is set at 0.0001×2^{Ls} . The \hat{df}_λ of the solution $\hat{\theta}_\lambda$ is the number of nonzero elements in the solution (the cutoff value for zero is 10^{-6}). The selected λ is the minimizer of the $\text{AIC}(\lambda)$ or $\text{BIC}(\lambda)$ over the grids of s and all the τ values which gives the LARS/Lasso solution along its path.

A. Experimental Setup

Our experimental setting is similar to the one in [1]. Five testing functions, namely, Blocks, Bumps, Blip, Corner, and Wave, are used that represent a variety of function characteristics. The test functions have been scaled so they all have a standard deviation of 1. Two levels of signal-to-noise ratio (SNR) are chosen: $\text{SNR} = 3$ and $\text{SNR} = 7$ (low and high), where SNR is defined as $\text{SNR} = \text{std}(\mathbf{f})/\sigma$. Two different sample size $n = 100$ and $n = 200$ are used. Nonequispaced sample points are generated by sampling with replacement from the grid $\{i/m : i = 1, 2, \dots, m\}$ with $m = 1024$.

For each of the above experimental factor combinations, 50 data sets are generated. For each simulated data set, our method and WK method of [1] are applied to estimate the test signal. Some tuning parameters for the WK method are: number of groups in which input data are split for cross validation is set at the default value 4, no blocking of the wavelet coefficients, all other parameters are set at the default values of the program. The root mean square error (RMSE) of the solution $\hat{\mathbf{f}}$ is recorded to assess its quality by its l_2 norm distance from \mathbf{f} , the true signal at the nonuniformly sampled points, $\text{RMSE} = \sqrt{\|\hat{\mathbf{f}} - \mathbf{f}\|^2/n}$.

B. Results

Boxplots of the RMSEs of the estimates obtained by our method and WK method are presented in Figs. 1–3 for Bumps, Blip, and Wave, respectively. The boxplots for Blocks and Corner are not presented here due to space constraints.

The x-axis labels in the box plot read as follows: best denotes the best estimates obtained by solving (1) for the λ in the search grid given the \mathbf{f} ; AIC denotes the estimates obtained by solving (1) with λ chosen by AIC using the estimated df (the number of nonzero elements in the solution of θ), WK standards for the wavelet kernel method of [1]. In almost all the cases, AIC outperforms BIC, so results by BIC are not plotted here.

For Bumps signal, AIC outperforms WK by a large margin and AIC is close to best overall. For Blip, AIC is comparable to WK. For Wave, WK outperforms AIC, and WK is close to best. Tests in Blocks signal not plotted here reveal that AIC outperforms WK. For Corner, WK slightly outperforms AIC. These

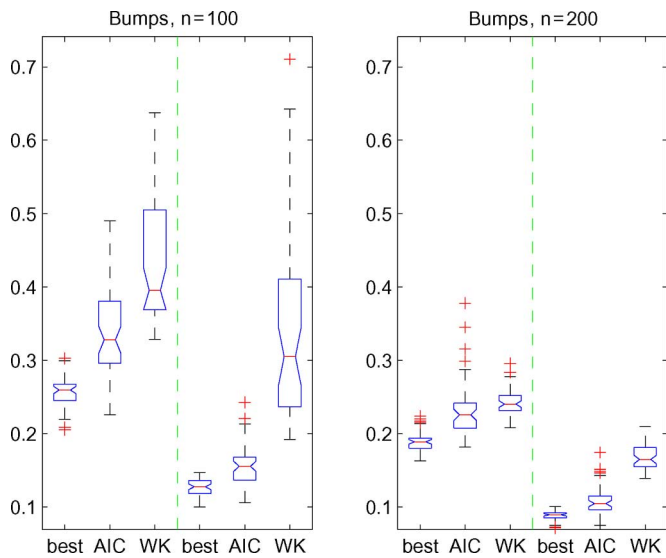


Fig. 1. Boxplots of the RMSEs for bumps signal. SNR = 3 for the left and SNR = 7 for the right panel within each subplot.

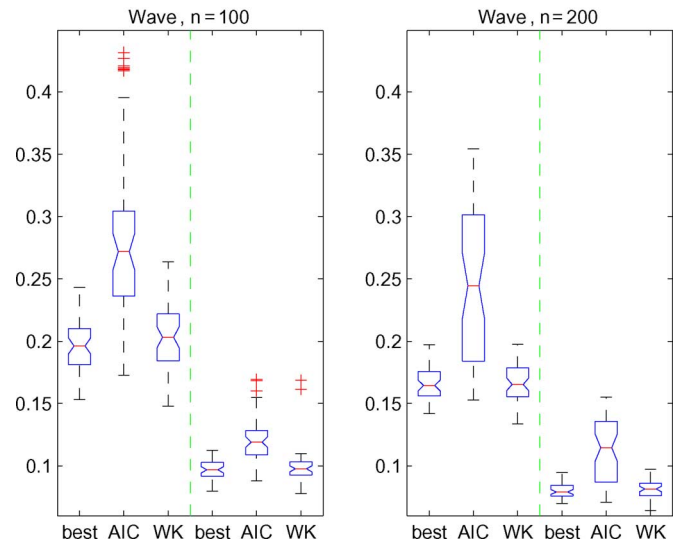


Fig. 3. Boxplots of the RMSEs for wave signal. SNR = 3 for the left and SNR = 7 for the right panel within each subplot.

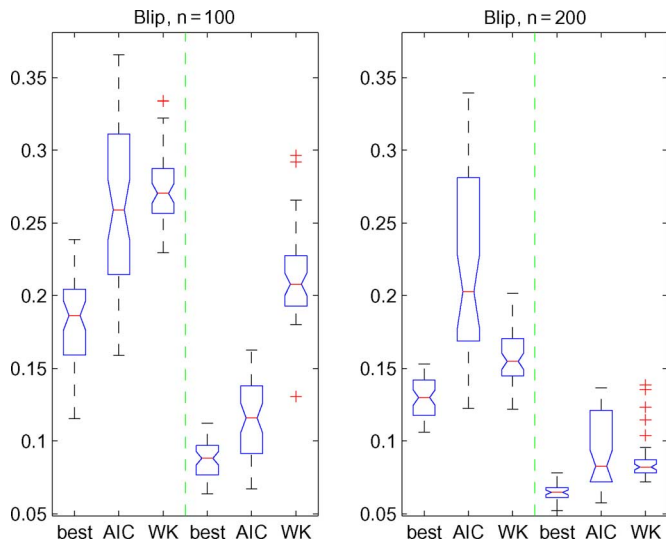


Fig. 2. Boxplots of the RMSEs for blip signal. SNR = 3 for the left and SNR = 7 for the right panel within each subplot.

results seem to suggest that AIC performs better in bumpier signals, and WK performs better in smoother signals. Overall, AIC performs well.

VI. CONCLUSION

In this letter, we propose a level dependent l_1 penalized wavelet reconstruction method for nonuniformly sampled signals. The LARS/Lasso algorithm is applied to solve the Lasso problem. The data adaptive choice of the two regularization parameters is based on the AIC, and the degrees of freedom is estimated by the number of nonzero elements in the Lasso solution. The empirical performance of the method is evaluated on the commonly used 1-D test signals. Results from these experiments illustrate that the proposed method possesses good empirical properties. Some future efforts include testing this

method on images and using nonorthogonal transforms such as translation invariant wavelet transform or curvelet transform.

REFERENCES

- [1] U. Amato, A. Antoniadis, and P. Marianna, "Wavelet kernel penalized estimation for non-equispaced design regression," *Statist. Comput.*, vol. 16, pp. 37–55, 2006.
- [2] A. Antoniadis and J. Fan, "Regularization of wavelet approximations," *J. Amer. Statist. Assoc.*, vol. 96, pp. 939–955, 2001.
- [3] H. Choi and R. G. Baraniuk, "Interpolation and denoising of nonuniformly sampled data using wavelet-domain processing," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1999, pp. 1645–1648.
- [4] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, pp. 1200–1224, 1995.
- [5] D. L. Donoho and Y. Tsaig, "Fast solution of l_1 -norm minimization problems when the solution may be sparse," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.
- [6] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, pp. 407–499, 2004.
- [7] M. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, Dec. 2007.
- [8] W. Fu, "Nonlinear GCV and quasi-GCV for shrinkage models," *J. Statist. Plan. Infer.*, vol. 131, pp. 333–347, 2005.
- [9] G. Hennenfent, E. van den Berg, M. P. Friedlander, and F. J. Herrmann, "New insights into one-norm solvers from the Pareto curve," *Geophysics*, vol. 73, no. 4, pp. A23–A26, 2008.
- [10] S. Mallat, *A Wavelet Tour of Signal Processing*. New York: Academic, 1998.
- [11] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. R. Stat. Soc. Ser. B*, vol. 58, pp. 267–288, 1996.
- [12] H. H. Zhang, G. Wahba, Y. Lin, M. Voelker, M. Ferris, R. Klein, and B. Klein, "Variable selection and model building via likelihood basis pursuit," *J. Amer. Statist. Assoc.*, vol. 99, pp. 659–672, 2004.
- [13] H. Zou, T. Hastie, and R. Tibshirani, "On the 'Degrees of Freedom' of the Lasso," *Ann. Statist.*, vol. 35, pp. 2173–2192, 2007.
- [14] B. Vidakovic, *Statistical Modeling by Wavelets*. New York: Wiley, 1999.
- [15] Y. Yang, "Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation," *Biometrika*, vol. 92, pp. 937–950, 2005.