

Ridgelet Pursuit : Application to Regression Estimation

Alain Rakotomamonjy

P.S.I INSA de Rouen,
Avenue de l'université
76801 Saint Etienne du Rouvray France
alain.rakotomamonjy@insa-rouen.fr

Abstract. This paper deals with a method for extending ridgelet based approximation to high-dimensional problem. Ridgelet approximation can be interpreted as a neural networks with a special neural activation function that has to satisfy an admissibility condition. This latter is similar to the one used in wavelet theory. Using such a function allows to have a set a function that is a frame of $L^2(\mathbb{R}^d)$. The algorithm used for approximation is based on a matching pursuit in a infinite dictionary that leads to a convergent algorithm. The selection of the best function is based on a Levenberg-Marquadt minimization algorithm. The exposed method is then applied to artificial regression problem on function with singularity and a financial dataset.

1 Introduction

Wavelet theory has become a widely used tool in signal and image processing for either approximation, estimation or compression [6]. However, due to the practical difficulty for building high-dimensional orthonormal wavelet basis, applications with dimensions higher than 2 are not frequent. The cooperation of neural networks framework and the wavelet theory allow to overcome this problem and then high-dimensional wavelet can be practically implemented [12]. The so-called wavelet network has emerged in the early nineties and combines the advantages of the wavelet with the learning process of a neural networks [13].

More recently, a new approach for wedding neural networks and wavelet has been proposed by Candès within the context of harmonic analysis [2]. The underlying idea is to look for neural activation function that after dilation and translation forms a set of $L^2(\mathbb{R}^d)$ frame. The condition for having such an activation function is similar to the one for wavelet :

$$\int \frac{|\hat{\psi}(\omega)|^2}{|\omega|^d} d\omega < \infty$$

where $\hat{\psi}$ is the Fourier Transform of the activation function. For now on, a function satisfying this condition is called a ridgelet. The advantage of frame is

that one can have a stable representation of any function in $L^2(\mathbb{R}^d)$ from the following expression :

$$f(x) = \sum_{i \in \Gamma} \langle f, \bar{\psi}_i \rangle \psi_i(x) \quad (1)$$

where $\{\psi_i\}_{i \in \Gamma}$ and $\{\bar{\psi}_i\}_{i \in \Gamma}$ are respectively a frame and dual frame of $L^2(\mathbb{R}^d)$. In an other paper [4, 3], Candès has shown that ridgelet-based function estimation can be theoretically highly efficient in high-dimensional problems . However, in this case, constructing a frame is prohibitive as the number of frame elements is exponentially related to the dimension of the problem. Hence, building explicitly the frame basis is not very appropriate. To overcome this problem, we propose a method for approximation, based on an adapted matching pursuit algorithm, that does not need the explicit knowledge of the frame elements, and we apply this algorithm to regression problems. This allow to exploit the information compression, robustness and localization properties of ridgelet.

In this paper, we recall briefly the ridgelet frame theory and describe the modified matching pursuit algorithm. Then, examples of application on artificial data and a real financial regression problem are given.

2 Ridgelet Approximation

Frame theory has been developed mainly for wavelet. However, it is a general concept that can also be extended for multidimensional case. Frames [5] have interesting properties that can be exploited for function approximations. So, in the first part of this section, we recall a brief theory about ridgelet, before describing the ridgelet pursuit algorithm.

2.1 Ridgelet Frames

A frame is a set of functions $\{\psi_i\}_{i \in \Gamma}$ of $L^2(\mathbb{R}^d)$ that satisfy the following condition :

$$A \|f\|_{L^2(\mathbb{R}^d)}^2 \leq \sum_{i \in \Gamma} |\langle f, \psi_i \rangle|^2 \leq B \|f\|_{L^2(\mathbb{R}^d)}^2 \quad (2)$$

with $0 < A \leq B < \infty$, for all function f in $L^2(\mathbb{R}^d)$. A dual frame $\{\bar{\psi}_i\}_{i \in \Gamma}$ can also be defined and by means of those two sets of functions, one can expanded a function f as :

$$f(x) = \sum_{i \in \Gamma} \langle f, \bar{\psi}_i \rangle \psi_i(x) = \sum_{i \in \Gamma} \langle f, \psi_i \rangle \bar{\psi}_i(x)$$

Frames can be considered as a redundant basis which redundancy are statistically useful [5, 11], as the more the frame elements is redundant, the more the reconstruction is robust to noise.

Ridgelet has been introduced by Candès whose aim was to propose a stable function representation with a sum of neural activation function [2]:

$$f(x) = \sum_i a_i \sigma(u_i^t x + b_i)$$

where u is a unit vector of \mathbb{R}^d , a_i and b_i two constants. As defined by Candes, ridgelet can be viewed as multidimensional projective wavelet. In fact, a frame of ridgelet ψ_γ is a set of functions obtained from the dilation, translation and rotation of a single ridgelet :

$$\psi_\gamma(x) = a_0^{j/2} \psi \left(a_0^j u^t x - kb_0 \right)$$

with $\gamma \in \Gamma_d = \left\{ (a_0^j, u, kb_0 a_0^{-j}), j \geq j_0, u \in \Sigma_j, k \in Z \right\}$, where Σ_j is a discretization of the unit sphere S^{d-1} of \mathbb{R}^d . This set is a frame of $L^2([0, 1]^d)$, however, this is not very restrictive as a simple renormalization allows to be respect this hypothesis. The main drawback of this frame is that its size N_j for a given dilation j is exponentially related to the dilation j considered, and the dimension of the problem. In fact, $N_j \propto a_0^{(j-j_0)(d-1)}$ and thus, for dimension higher than 2, using equation (1) is prohibitive, because it needs explicitly the construction of the frame and dual frame elements.

This is the main justification of the algorithm that we propose in the next part.

2.2 Ridgelet Pursuit

Our aim is to approximate any function $f \in L^2([0, 1]^d)$ with a linear combination of ridgelet :

$$f(x) = \sum_{\gamma \in \Gamma} \omega_\gamma \psi_\gamma$$

with $\omega_\gamma \in \mathbb{R}$ and Γ is a set of index that we will define later. The approximation should not need any construction of the frame elements and should be convergent in the sense :

$$\left\| f - \sum_{n=1}^N \omega_n \psi_n \right\| \leq C^{-N} \|f\| \quad \text{with } 0 \leq C < 1 \quad (3)$$

The dictionary of ridgelet is composed of vectors of unitary norm and it must include N vectors that form a frame of $L^2([0, 1]^d)$. For instance, the dictionary D can be the set of ridgelet :

$$\psi_\gamma \text{ with } \gamma \in \Gamma_d = \left\{ (a_0^j, u, kb_0 a_0^{-j}), j \geq j_0, u \in S^{d-1}, k \in Z \right\}$$

as this set satisfies the above conditions.

The matching pursuit algorithm consists in approximating a function f with elements of a dictionary D . At first, the algorithm looks for a vector g_{γ_0} which minimize the norm of a residual defined as :

$$f = \omega_{\gamma_0} g_{\gamma_0} + Rf$$

Then, one has to reiterate this algorithm on the residual and so on. Finally, after m iterations, the approximation \hat{f} can be written as :

$$\hat{f} = \sum_{m=0}^{M-1} \langle R_m f, g_{\gamma_m} \rangle + R_{m+1} f$$

where $R_0 f = f$ and $R_m f$ is the residual after $m - 1$ iterations (for more details, readers should refer to [8, 6]). In our case, this matching pursuit algorithm can not be applied as it is, because, the ridgelet dictionary is not defined explicitly and is of infinite size.

Recall that at each step, the aim is to find a ridgelet that minimize $\|R_{m+1} f\|^2$ or $\|R_m f - \langle R_m f, g_{\gamma_m} \rangle g_{\gamma_m}\|^2$ which is equivalent to find a ridgelet that maximize $|\langle R_m f, g_{\gamma_m} \rangle|$. As this search may be time-consuming, it is sufficient to find a sub-optimal ridgelet that satisfies, for a given α :

$$|\langle R_m f, g_{\gamma_m} \rangle| \geq \alpha \sup_{\gamma \in \Gamma_d} |\langle R_m f, g_{\gamma} \rangle| \quad (4)$$

with $0 < \alpha \leq 1$. As the set of ridgelet is infinite, an iterative method is used for looking for the optimal ridgelet that minimize the norm of the residual. Here, the problem is to minimize a quadratic cost with regards to the parameters γ , namely, j , k and u . In this case, for instance, a Levenberg-Marquadt algorithm [1, 7] can be used for solving this problem. It can be shown that [9], there exists a constant C , depending on the frame bounds, the cardinality of the frame elements included in the dictionary, and the suboptimal parameter α , so that for every $m > 0$, equation (5) holds.

Theorem 1. *it exists a constant depending on the frame bound and the suboptimal parameter α so that for all $m \geq 0$*

$$\|R^m f\| \leq C^m \|f\| \quad (5)$$

and so

$$f = \sum_{m=0}^{+\infty} \langle R^m f, g_{\gamma_m} \rangle g_{\gamma_m} \quad (6)$$

and

$$\|f\|^2 = \sum_{m=0}^{+\infty} |\langle R^m f, g_{\gamma_m} \rangle|^2 \quad (7)$$

2.3 The Ridgelet Pursuit algorithm

The algorithm can be summarized as follows :

1. At first, one has to initialize the residual with the data to be approximated : $m = 0$ and $R_m f = f$.

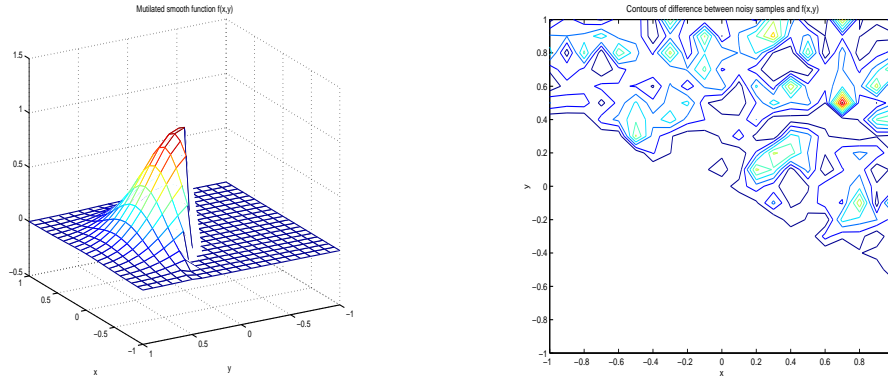


Fig. 1. The original mutilated function and the contour of L_1 difference between ideal and noisy samples.

2. Look for the best function g_{γ_m} that minimize the norm of the residual. This is done with a gradient-descent algorithm into the parameter space of the ridgelet dictionary (usually, these parameters are the dilation, translation and direction of ridgelets).
3. Update the residual.
4. Verify if the stopping criteria is satisfied unless go back to step 2. Commonly, this criteria can be the number of ridgelet in the model, the mean-square error for the learning set, or the mean-square error in a validation set.

3 Application to Regression Estimation

This algorithm of ridgelet pursuit has been tested on regression problems.

3.1 Example 1 : 2D Regression on noisy mutilated function

First example consists in the approximation of a noisy smoothed mutilated function $f(x, y)$:

$$f(x, y) = 1_{\{x+2y>0\}} e^{-4(x^2+y^2)}$$

This function has been regularly sampled on the interval $[-1, 1]^2$, then a gaussian noise with a standard deviation of 0.15 is added to the samples. Two parameters control the complexity of the model, the first one, is the number or ridgelet used for estimating f . In fact, according to the property of convergence of the algorithm, the approximation \hat{f} will interpolate the noisy data if too many ridgelets are used. The second complexity parameters is the number of dilation j explored in the ridgelet dictionary. As this dilation parameter is related to the frequency contents of the ridgelet, hence, removing high dilation can be interpreted as a frequency filtering, and thus, is equivalent to approximate data with smoother functions. In figure (2) and (3), typical results are depicted. We have

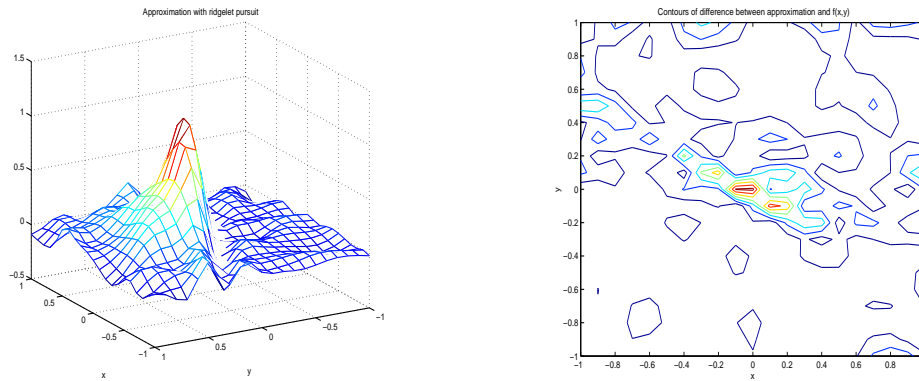


Fig. 2. approximation with 25 ridgelet and the contour of L_1 difference between ideal and approximation function.

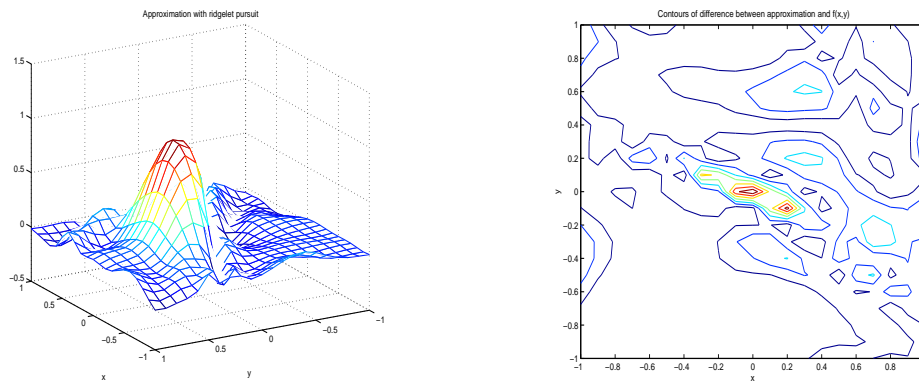


Fig. 3. approximation with 15 ridgelet and the contour of L_1 difference between ideal and approximation function.

the estimation of the regression function and the contour of the L_1 error. One can note that the discontinuity is well approximated by the ridgelet expansion and smooth regions have been partially denoised for both models.

3.2 Example 2: Regression on financial dataset

This second example comes from the NIPS learning competition datasets and consists in the prediction of the value of one asset based on the knowledge of 5 others. The data set is composed of 200 examples for learning (which have been separated in 150 for learning and 50 for model selection) and 200 others for testing. Dilations explored have been fixed to 2^j with $j \in \{1 \dots 4\}$ and the complexity of the approximation is controlled by the number of ridgelet.

Figure (4) depicts the predicted value with regards to the real value for a model with 20 ridgelets and the prediction of the asset values on the testing set

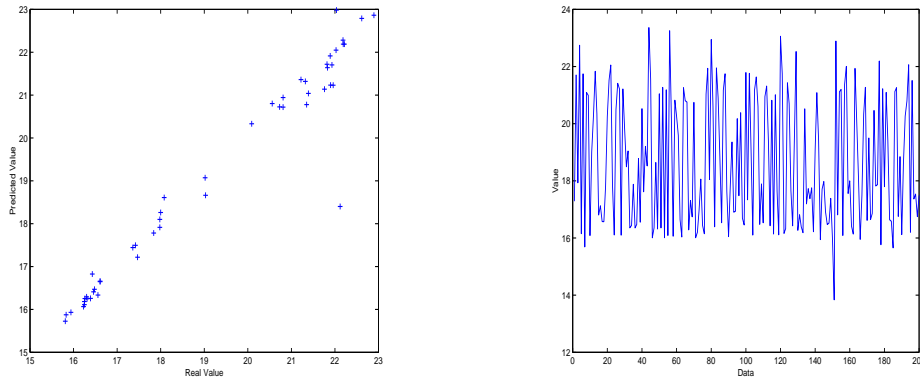


Fig. 4. left : Predicted value of asset vs real value for the validation set. Right : Prediction of asset value for the testing set

for the same. One can notice that in this 5-dimensional problem, the algorithm behaves well. Outliers can also be noted in the both the validation and testing set.

4 Conclusions

A ridgelet pursuit algorithm have been proposed in this paper, and the examples given in here show promising results. The behaviour of this approximation method allows to extend the application of ridgelet to higher dimension than 2 with still good results, as one has seen in the second example. Besides, owing to the ridgelet frame properties of redundancy, compression and localization, this algorithm allows good approximation in region with discontinuity. Other aspects like the choice of dilation and other complexity control methods (penalisation...), variance of the model (due to the Levenberg-Marquardt algorithm) have to be investigated.

5 Appendix

The financial data can be found on the following URL :
<http://q.cis.uoguelph.ca/~skremer/NIPS2000/>

References

1. C. Bishop. *Neural Networks for Pattern Recognition*. Oxford Univ. Press, 1995.
2. E. Candès. Harmonic analysis of neural network. *Applied and Computational Harmonic Analysis*, (6):197–218, 1999.
3. E. Candès. On the representation of sobolev functions. Technical report, Stanford University, 1999.

4. E. Candès. Ridgelets : estimating with ridge functions. Technical report, Stanford University, 1999.
5. I. Daubechies. *Ten Lectures on Wavelet*. SIAM, 1992.
6. S. Mallat. *A wavelet tour of signal processing*. Academic Press, 1998.
7. Mathworks. *Optimization toolbox User's Guide*. The Mathworks Inc., 1997.
8. S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Trans Signal Processing*, 41(12):3397–3415, 1993.
9. A. Rakotomamonjy. Ridgelet Pursuit. Technical Report (in French), PSI AR-2000, 2000.
10. S. Soltani, S. Canu, and D. Boichu. A regression estimation method based on wavelet frames. In *ICANN*, 1998.
11. S. Soltani, Application de la transformée en ondelettes en reconnaissances de formes Ph.D. thesis (In french), Univ. Tech. Compiègne, 1999.
12. Q. Zhang and A. Benveniste. Wavelet networks. *IEEE Trans. on Neural Networks*, 3(6):889–898, 1992.
13. Q. Zhang. Using wavelet network in non parametric estimation. Technical report, IRISA, 1993.