Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Exploração Estatística do *Compressive Sensing* Utilizando Dados Sísmicos

Éberton da Silva Marinho[1]

Programa de Pós-Graduação em Ciências e Engenharia do Petróleo, UFRN, Natal, Brasil

Francisco Iranildo Ferreira do Nascimento Gomes[2]

Programa de Pós-Graduação em Ciências e Engenharia do Petróleo, UFRN, Natal, Brasil

Gilberto Corso[3]

Departamento de Biofísica e Farmacologia, Centro de Biociências, UFRN, Natal, Brasil

Liacir dos Santos Lucena[4]

Departamento de Física Teórica e Experimental, UFRN, Natal, Brasil

International Center for Complex Systems, UFRN, Natal, Brasil

Programa de Pós-Graduação em Ciência e Engenharia do Petróleo, UFRN, Natal, Brasil

Resumo. Computamos a correlação entre a taxa de erro do sinal recuperado pelo *Compressive Sensing* e os valores estatísticos: coeficiente de variação, assimetria e curtose; para um sismograma. Utilizamos os algoritmos de recuperação: *Bayesian Compressive Sensing*, l_1-*MAGIC* e StOMP CFAR. Obtivemos fortes evidências de correlações dos dados recuperados com a Curtose.

Palavras-chave. *Bayesian Compressive Sensing*, *Stagewise Orthogonal Matching Pursuit* (StOMP CFAR), l_1-*MAGIC*, *Wavelets*, Curtose.

1 Introdução

O teorema da amostragem de *Shannon-Nyquist* afirma que para restaurar um sinal exata e unicamente é necessário tê-lo amostrado no mínimo com o dobro de sua freqüência. Porém, a utilização do teorema de *Shannon-Nyquist* faz com que a quantidade de informação adquirida seja imensa, o que pode tornar intratável a análise e a manipulação de tanta informação pelos sistemas de informação atuais. Grandes volumes de dados são

[1] ebertonsm@gmail.com, eberton.marinho@ifrn.edu.br
[2] ironrdc@gmail.com
[3] gfcorso@gmail.com
[4] liacir.lucena@gmail.com

frequentemente redundantes e métodos para comprimí-los exercem um papel cada vez mais importante nas mais diversas aplicações [3].

Uma transofrmação muito utilizada atualmente para comprimir dados têm sido as transformadas *Wavelets* [7], com as quais os sinais são representados em termos de objetos matemáticos elementares, também chamados de "átomos", localizados no tempo e frequência. As propriedades de localização destes elementos de tempo-frequência levam a representações compactas de muitos sinais naturais [7]. No entanto, as técnicas tradicionais de compressão de dados utilizando *Wavelets* requerem toda a informação do sinal para realizar sua compressão [2, 3].

A teoria do *Compressive Sensing* (CS) [2, 3] tem demonstrado que se o sinal for esparso em alguma base e houver uma matriz de sensoriamento com algumas propriedades especiais, então, com um relativo pequeno número de projeções adequadas, conseguimos recuperar o sinal original com um baixo limiar de erro. Desta forma, a CS não é apenas um método de compressão de dados, mas sim de um processo de redução do volume da aquisição e recuperação de dados [2, 3].

Apesar de haver pesquisas utilizando a Transformada *Wavelet* Discreta (*Discrete Wavelet Transform* - DWT) em CS [12], a utilização desta técnica na área do petróleo é recente. Neste trabalho, temos como principal objetivo correlacionar as estatísticas: coeficiente de variação, assimetria e curtose; de traços sísmicos, com o erro da recuperação pelo CS. As técnicas de CS utilizadas foram: *Bayesian Compressive Sensing* [5], l1-*MAGIC* [1] e StOMP CFAR [4]. Em nossos testes, foram empregadas taxas de amostragem referentes a 10%, 20%, 30%, 40% e 50% do sinal original e, além disso, utilizamos como bases as *Wavelets* discretas: *Biorthogonal* 3.9, *Daubechies* 6, *Coiflet* 3 e *Symlet* 9.

2 Fundamentação Teórica

2.1 Os dados sísmicos

Entre as diversas técnicas de levantamento geofísico, a reflexão sísmica é uma das mais utilizadas e conhecidas. Cada traço sísmico é o resultado da captação das informações refletidas pelo sistema de camadas terrestres e a junção desses traços sísmicos forma um sismograma [9]. A imagem que trabalhamos é composta por 33 traços sísmicos. Nosso trabalho foi realizado utilizando a ferramenta de programação MATLAB [8]. Na Figura 1, em (a), temos um sismograma onde no eixo horizontal é representado a distância da fonte (pés) a cada um dos 33 geofones. No ponto 0, encontra-se o local da explosão. Ainda na Figura 1 (a), no eixo vertical temos o tempo em segundos a partir do momento da explosão. Na Figura 1, em (b), temos o sinal original (em azul), e em (c), a comparação com o sinal recuperado (em vermelho) pelo método l_1-*MAGIC*.

2.2 *Compressive Sensing*

Seja um sinal $x[i]$ de elementos, $i = 1,\ 2,\ ...,\ N$. Utilizando uma matriz base $\Psi = [\psi_1 \mid \psi_2 \mid ... \mid \psi_N]$ $N \times N$ construída numa base ortornormal (em nosso trabalho foram utilizadas bases *Wavelets* [12]). Neste caso, x pode ser expresso como

$$x = \sum_{i=1}^{N} s_i \psi_i \qquad \text{ou} \qquad x = \Psi\, s \tag{1}$$

onde s é um vetor coluna de coeficientes ponderados $s_i = \langle x, \psi_i \rangle = \psi_i^T x$. Desta forma, s e x são representações equivalentes do sinal, com x no domínio do tempo e s no domínio de Ψ. O sinal x é K-esparso se ele puder ser escrito como uma combinação linear de apenas K vetores da base, $K << N$.

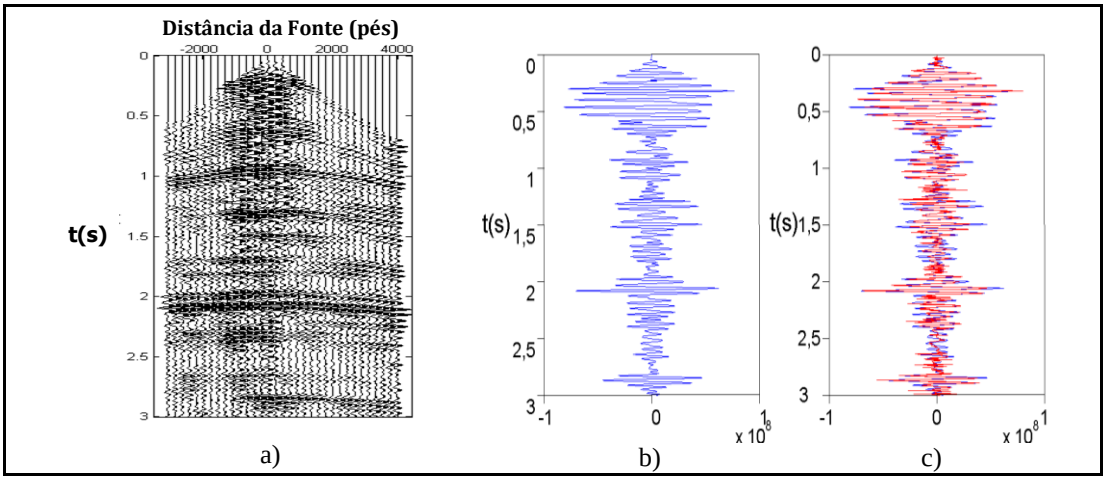

Figura 1: Em (a), sismograma composto de 33 traços de um perfil geológico. Em (b), traço sísmico original de número 16 com 1501 pontos, e em (c), superposição do traço sísmico original (em azul) e o recuperado (em vermelho) utilizando o método l_1-*MAGIC*, com 30% de amostras e a *Wavelet Coiflet* 3.

Atualmente, as técnicas de aquisição e compressão de informação sofrem de três ineficiências inerentes: O número N de aquisições pode ser muito grande em comparação ao número de esparsidade K; um número N de coeficientes deve ser computado mesmo que $(K - N)$ coeficientes sejam jogados fora depois da transformação; a localização dos coeficientes significativos deve ser codificada, o que introduz um atraso. A CS soluciona esta ineficiência adquirindo diretamente um sinal comprimido sem passar pelo estágio intermediário de adquirir N amostras [2, 3]. Considere um processo de medição linear geral $\Phi = \{\phi_j\}_{j=1}^{M}$ que computa $M < N$ produtos internos entre x e a coleção de vetores como em $y_j = \langle x, \phi_j \rangle$. Pela substituição em (1), podemos escrever

$$y = \Phi\ x = \Phi\ \Psi\ s = \Theta\ s \tag{2}$$

onde $\Theta = \Phi\ \Psi$ é uma matriz $M \times N$. Ressaltamos que o processo de medição é não adaptativo, significando que Θ não depende de x. O problema consiste em desenvolver: i) uma matriz de medição estável Θ tal que a informação relevante em qualquer sinal K-

esparso ou compressível não seja prejudicado pela redução de dimensionalidade de x ; ii) um algoritmo de recuperação que reconstrua o x a partir de tão somente $M \sim K$ medições y .

2.2.1 *Bayesian Compressive Sensing*

O problema de recuperação pode ser formulado com uma abordagem Bayesiana, onde o *Relevance Vector Machine* (RVM) proposto em [13] é adaptado para o problema da CS. Na modelagem Bayesiana o sinal s desconhecido é associado a uma probabilidade $p(s \mid \gamma)$ (distribuição a priori), que modela nosso conhecimento sobre a natureza de s . As observações de y também são um processo aleatório com distribuição condicional $p(y \mid s, \beta)$, onde $\beta = 1/\sigma^2$ é o inverso da variância do ruído. Estas distribuições dependem do modelo de parâmetros γ e β , que são hiperparâmetros e distribuições a priori adicionais, chamadas hiperprioris.

2.2.2 l_1-*MAGIC*

Seja $s \in \Re^N$ um sinal esparso. Ele pode ser recuperado a partir de um pequeno número de medições lineares $y = \Theta s \in \Re^K$, $K << N$ ou $y = \Theta s + e$ pela resolução de um problema convexo [1, 2], onde e representa o erro. Tais problemas convexos podem ser encaixados em duas classes de problemas: aqueles que podem ser classificados como de programação linear (*Linear Programming*-LP); aqueles que podem ser classificados como de programação de cones de segunda ordem (*second-order cone programs*-SOCPs). No pacote l_1-*MAGIC*, Candès e Romberg [1] criaram diversos algoritmos para recuperação utilizando a norma l_1.

2.2.3 StOMP CFAR

O *Stagewise Orthogonal Matching Pursuit* (StOMP) é um algoritmo simples, para obtenção de solucões aproximadamente esparsas de certos sistemas indeterminados de equações lineares, que seleciona e projeta os limiares iterativamente [4]. O StOMP é um método guloso (seleciona sempre a melhor solução local na tentativa de encontrar a melhor solução global) e iterativo baseado no *Orthogonal Matching Pursuit* (OMP) [10]. Porém, enquanto o OMP seleciona apenas um coeficiente a cada iteração, o StOMP seleciona vários coeficientes a cada repetição do algoritmo. Há dois tipos de StOMP: o CFAR e o CFDR [4]. Neste trabalho utilizamos o StOMP CFAR.

3 Análises Estatísticas

Fizemos a recuperação do sinal utilizando 10, 20, 30, 40 e 50% dos pontos de cada um dos 33 traços sísmicos e medimos a taxa de erro para os métodos BCS, l_1-*MAGIC* e StOMP CFAR. Considerando que o sinal original é um vetor representado por x e o sinal recuperado por x', a taxa de erro relativa do sinal é $\| x - x' \|_2 / \| x \|_2$. Através de uma matriz

DOI: 10.5540/03.2016.004.01.0044

aleatória gaussiana, com média 0 e desvio padrão 1/N, sensoriamos os coeficientes com as seguintes *Wavelets*: *Biorthogonal* 3.9; *Daubechies* 6; *Coiflet* 3 e *Symlet* 9. Para cada conjunto de parâmetros citados anteriormente executamos 100 testes e trabalhamos com a média aritmética da taxa de erro.

Para a execução dos testes pelo método BCS utilizamos a função *BCS_fast_rvm*, enquanto que para o método StOMP CFAR, utilizamos a função *SolveStOMP*. Ambos inseridos na biblioteca SparseLab versão 2.1 [11]. Por sua vez, para o l_1-*MAGIC*, utilizamos a função *l1eq_pd* da biblioteca l1magic versão 1.11 [6].

O coeficiente de variação, a assimetria e a curtose foram computados para cada um dos 33 traços e correlacionados à taxa de erro da recuperação. Ao todo foram calculadas 180 correlações (3 métodos de recuperação, 5 taxas de amostragem, 4 *Wavelets*, 3 estatísticas) e gerados gráficos para os pares ordenados (estatística versus taxa de erro). Para cada gráfico, determinamos os parâmetros dos estimadores para o modelo de regressão linear.

A Figura 2 ilustra um estudo de como foi compilado o coeficiente de correlação ρ, o valor de t_0 e o valor de P. Neste exemplo, apresentamos a taxa de erro versus o coeficiente de variação para os 33 traços do nosso sismograma utilizando como método de recuperação o l_1-*MAGIC*, com 30% das amostras do sinal original e usando a base *Wavelet Coiflet* 3.

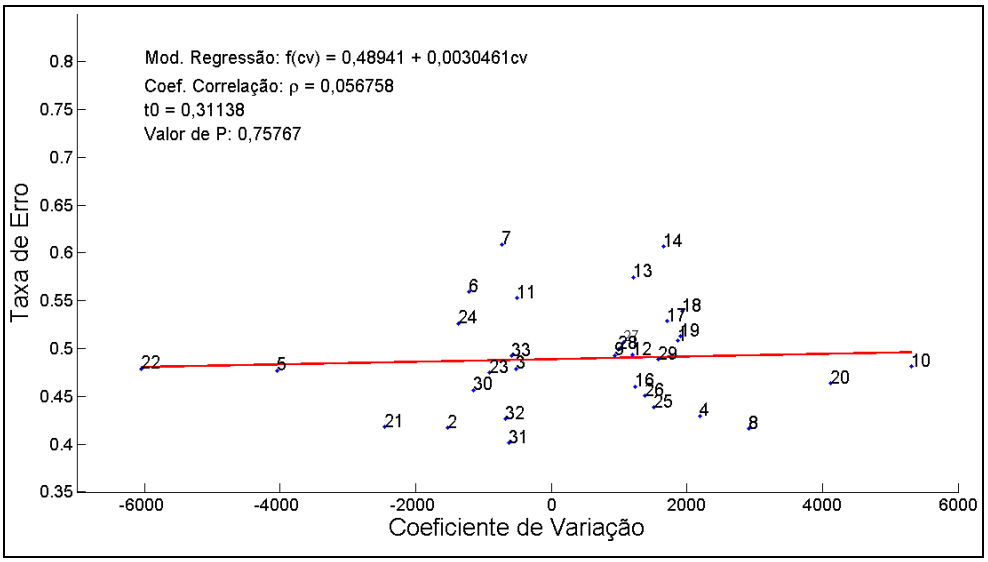

Figura 2: Correlação entre o coeficiente de variação e taxa de erro para os 33 traços sísmicos. Os valores do coeficiente de correlação e de P se encontram na legenda desta figura.

Computamos os valores de P das correlações e os organizamos na Tabela 1. Resultados significativos (P < 0,05) são destacados em negrito. Por questões didáticas, apresentamos os valores de P apenas das amostras 10, 30 e 50%.

Com base nos dados completos da Tabela 1, apresentamos na Tabela 2 a contagem da quantidade de testes significativos (P < 0,05) para cada método de recuperação, utilizando as três estatísticas. Os resultados obtidos nos levam a crer que a taxa de erro não tem correlação com a assimetria e o coeficiente de variação, mas sim com a curtose.

Tabela 1: Resultados para o valor de P usando os método BCS, l_1-*MAGIC* e StOMP CFAR

	% das amos-tras	BCS				l_1-MAGIC				StOMP CFAR			
		Wavelets				Wavelets				Wavelets			
		Bior3.9	Coif3	Db6	Sym9	Bior3.9	Coif3	Db6	Sym9	Bior3.9	Coif3	Db6	Sym9
Coef. Var.	10%	0,254	0,686	0,702	0,419	0,414	0,644	0,507	0,651	0,092	0,282	0,231	0,711
	30%	0,167	0,548	0,480	0,254	0,458	0,757	0,512	0,418	0,084	0,275	0,545	0,772
	50%	0,132	0,679	0,387	0,461	0,114	0,839	0,401	0,241	0,090	0,973	0,195	**0,021**
Ass.	10%	0,231	0,221	0,668	0,480	0,168	0,228	0,520	0,405	0,556	0,654	0,299	0,321
	30%	0,619	0,569	0,301	0,529	0,338	0,345	0,316	0,432	0,864	0,260	0,456	0,692
	50%	0,798	0,691	0,481	0,848	0,491	0,418	0,280	0,385	0,567	0,406	0,245	0,381
Curt.	10%	**0,007**	**0,001**	**0,001**	**0,001**	**0,002**	**0,003**	**0,001**	**0,001**	0,501	0,657	0,888	0,131
	30%	**0,015**	**0,014**	**0,037**	**0,014**	**0,001**	**0,003**	**0,005**	**0,001**	0,332	0,210	**0,001**	0,296
	50%	**0,038**	0,072	**0,042**	**0,048**	**0,002**	**0,009**	**0,005**	**0,001**	**0,024**	**0,030**	**0,008**	**0,008**

Tabela 2: Quantidade de testes significativos, de um total de 33 traços (P<0,05)

	Métodos de recuperação do sinal		
	BCS	l_1-*MAGIC*	StOMP CFAR
Coef. de Variação	0	0	2
Assimetria	0	0	0
Curtose	18	20	10

4 Conclusões e Trabalhos Futuros

Este trabalho aplica a técnica de CS a 33 traços sísmicos utilizando vários métodos de recuperação, porcentagens de amostragem e bases *Wavelets*. Comparamos a taxa de erro de recuperação com os valores estatísticos: coeficiente de variação, assimetria e curtose do dado original. Obtivemos fortes indícios que há uma correlação entre a taxa de erro e a curtose nos dados sísmicos.

O fato da taxa de erro do sinal recuperado estar correlacionada com a curtose do sinal, parece indicar que quanto mais o sinal estiver concentrado em torno da média, mais eficientemente será processado pelo CS. O fato do coeficiente de variação e a assimetria não se correlacionarem com a performance do CS é intrigante. No intuito de elucidar estas questões, temos como perspectivas futuras realizar análises similares para outras medidas estatísticas como a Entropia, Coeficiente de Gini, DFA, Expoente de Hurst, bem como utilizar um maior número de traços sísmicos a fim de explorar este universo de correlações.

DOI: 10.5540/03.2016.004.01.0044

Referências

[1] E. Candès and J. Romberg, l1-magic: Recovery of sparse signals via convex programming, vol.4, (2005), to appear in http://users.ece.gatech.edu/justin/l1magic/downloads/l1magic.pdf.

[2] E. Candès, J. Romberg and T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, IEEE Transactions on Information Theory, vol. 52, 489 - 509, (2006), DOI: 10.1109/TIT.2005.862083.

[3] D. Donoho, Compressed Sensing, IEEE Transactions on Information Theory, vol. 52, 1289 - 1306, (2006), DOI: 10.1109/TIT.2006.871582.

[4] D. L. Donoho, Y. Tsaig, I. Drori and J.-L. I. Starck, Sparse Solution of Underdetermined Systems of Linear Equations by Stagewise Orthogonal Matching Pursuit, IEEE Transactions on Information Theory, vol. 58, 1094-1121, (2012), DOI: 10.1109/TIT.2011.2173241.

[5] S. Ji, Y. Xue and L. Carin, Bayesian Compressive Sensing, Signal Processing, IEEE Transactions on, V. 56, 2346-2356, (2008), DOI: 10.1109/TSP.2007.914345

[6] ℓ_1-MAGIC. Disponível em: < http://users.ece.gatech.edu/~justin/l1magic>. Acesso em: 3 de maio de 2015.

[7] S. Mallat, A wavelet tour of signal processing, 2nd ed., Academic Press, (1998).

[8] MATLAB and Statistics Toolbox Release 2012b, The MathWorks, Inc., Natick, Massachusetts, United States, (2012).

[9] W. A. Mousa and A. A. Al-Shuhail, Processing of Seismic Reflection Data Using MATLAB, Synthesis Lectures on Signal Processing, Morgan & Claypool Publishers, Cap. 1 e 4, (2011).

[10] Y. C. Pati, R. Rezaiifar and P. S. Krishnaprasad, Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition, Signals, Systems and Computers, Conference Record of The Twenty-Seventh Asilomar Conference on, IEEE, vol. 1, 40-44, (1993), DOI: 10.1109/ACSSC.1993.342465.

[11] SparseLab. Disponível em: <https://sparselab.stanford.edu>. Acesso em: 3 de maio de 2015.

[12] J.-L. Starck, F. Murtagh and J. M. Fadili, Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity, Cambridge University Press, (2010).

[13] M. Tipping, Sparse Bayesian learning and the relevance vector machine, The Journal of Machine Learning Research, vol. 1, 211–244, (2001).

