MULTI-SCALE SPARSE CODING WITH ANOMALY DETECTION AND CLASSIFICATION

Hojjat Akhondi-Asl, James D. B. Nelson

Department of Statistical Science, University College London, UK
{h.akhondi-asl, j.nelson}@ucl.ac.uk

ABSTRACT

We here place a recent joint anomaly detection and classification approach based on sparse error coding methodology into multi-scale wavelet basis framework. The model is extended to incorporate an overcomplete wavelet basis into the dictionary matrix whereupon anomalies at specified multiple levels of scale are afforded equal importance. This enables, for example, subtle transient anomalies at finer scales to be detected which would otherwise be drowned out by coarser details and missed by the standard sparse coding techniques. Anomaly detection in power networks provides a motivating application and tests on a real-world data set corroborates the efficacy of the proposed model.

1. INTRODUCTION

Occam's razor principle states that if two models can explain an event, then the one that is more parsimonious is typically better and more robust. There is much ongoing interest in the use of sparse regularisation techniques to exploit the common phenomena that natural signals can be described as a sparse combination of basis elements either in their natural domain or some alternative transform domain. Sparse coding, the process that models data vectors as a sparse linear combination of basis components, has proven to be effective and robust for many applications such as signal reconstruction [1, 2] and classification [3] and is widely used in many fields, including signal processing, machine learning, and statistics.

Sparse coding with anomaly detection and convex optimisation has been recently considered in [4]. Anomaly detection is a problem that aims to detect outliers/faults/anomalies in the data that do not conform to an expected model. It is used in many applications such as fault detection in low-voltage data, data security, fraud detection, event detection in sensor networks, detecting Eco-system disturbances and others [5]. In [4], Adler et al. applied the K-SVD method to learn a dictionary from ECG data and applied sparse coding with anomaly detection to detect irregular heartbeats. In [4], the same approach was used to remove specular reflectance and shadows from a set of images. In [6], Kalaitzis et al. considered a similar approach to detect and classify faults/anomalies in a three-phase low-voltage time series data. In both papers [4, 6], the sparse coding idea is used to balance error against sparsity of the solution, thereby allowing one to simultaneously detect anomalies and represent data vectors with a sparse representation.

A drawback of the standard sparse error-coding framework is that anomalies at all scales are treated the same. The anomaly detection schemes discussed above can be improved by incorporating the wavelet transform to better detect the early signs of failures in the data. The wavelet transform is a multi-scale time-frequency transform that with its good time-frequency resolution can accurately detect discontinuities and sudden changes in signals. This motivates us here to propose a multi-scale, wavelet-based sparse error-coding approach whereby an overcomplete wavelet basis is incorporated. Anomalies can then be more apparent at certain scale levels more clearly than they otherwise would be by the standard mode studies in [4, 6]. In this paper we propose a novel convex optimisation scheme for fault detection and classification using the multi-scale stationary wavelet transform. We place the approach into an alternating direction method of multipliers (ADMM) [7] framework and show, with a real-world voltage traces, that the method demonstrates utility not currently possible with the standard sparse-coding model.

The rest of the paper is organised as follows. In Section II, some background on sparse coding, dictionary design and stationary wavelet transform is presented. In Section III we frame the sparse coding for detecting anomalies in three-phase low-voltage time series data with multi-scale stationery transform. With an ADMM-based implementation we show that our method can accurately detect anomalies in the data which are not picked up by the standard sparse error-coding method. In Section IV, using a similar approach in Section III, we present our method that can simultaneously detect and classify faults using stationary wavelet transform. In Section V, we offer conclusions and ideas for possible future directions.

2. BACKGROUND

Consider a data vector $y \in \mathbb{R}^{n\times 1}$ that can be expressed as $y = Ds$ where $D \in \mathbb{R}^{n\times K}$ is called the dictionary, a database that contains the features/atoms as its columns. Here, $s \in \mathbb{R}^{K\times 1}$ is a vector that represents the signal y from the dictionary D.

H. Akhondi-Asl and J. D. B. Nelson are supported by EPSRC grant EP/N508470/1; in addition, J. D. B. Nelson is supported by grants from the Dstl.

When the dictionary is a basis, every signal can be uniquely represented as the linear combination of the dictionary atoms. If D is an overcomplete basis, where the dimension of the data vector is smaller than the number of atoms of the dictionary ($n \leq K$), then there are infinite number of solutions to the above model. However, by imposing sparsity on the vector s, it is possible to obtain the sparsest solution. Overcomplete dictionaries, with the extended feature vectors in the basis, have the ability of better capturing the structures and pattern in the data. To achieve sparsity, an additional term needs to be introduced on the vector s. The most direct measure of sparsity is the ℓ_0 norm on s, where $||s||_0$ is a quasi-norm which counts the number of non-zero elements in the vector s. The above discussion can be described with the following minimisation model:

$$\min_s \frac{1}{2}||y - Ds||_F^2 + \lambda||s||_0. \quad (1)$$

Here, the first term can be interpreted as a reconstruction or log-likelihhod term that forces the objective function to yield an acceptable representation of the vector s and the second term imposes sparsity on the vector s; it is equivalent to a log-prior on the solution. The penalty parameter λ is a regularisation parameter on the sparsity. The above minimisation problem is NP-hard, however there are many available pursuit algorithms, such as matching pursuit [8] and orthogonal matching pursuit [9], that can find an approximate solution to the problem. In practice however, it is common to apply a convex relaxation and use $||s||_1$ instead to impose sparsity via a Laplacian log-prior:

$$\min_s \frac{1}{2}||y - Ds||_F^2 + \lambda||s||_1. \quad (2)$$

Thus, the problem reduces to the LASSO [10] or the Basis Pursuit problem [11]. The ℓ_1 norm is convex and therefore the solution to the norm is unique. In the framework discussed in the introduction, the choice of dictionary is of utmost importance. Generally, dictionary design can be done in one of the following ways: (1) Analytically: By a mathematical model of the data such as transforming the signal to the Fourier domain, STFT, wavelet domain, etc. These methods lead to highly structured dictionaries with fast numerical implementations [12, 13]. (2) Data-Driven: By learning the dictionary from a set of training set. Examples include Principal Component Analysis [14], Generalised PCA [15], Method of Optimal Directions (MOD) [16] and the K-SVD method [2]. These methods lead to dictionaries that are much more adaptive to the data. (3) Combination of signal transformation and the data driven approach [17]. Our aim is to design a dictionary that can best describe the signal and that is most sensitive to faults. The dictionary can be designed with the actual low-voltage time series samples [6], however this approach lacks frequency resolution, which is necessary for detecting abrupt faults. Instead, we consider the incorporation of an overcomplete wavelet basis into the dictionary. Unlike [6], this results in a dictionary which can better model the healthy signal, resulting in an accurate detection of discontinues and sudden changes in the data. With the discrete wavelet transform (DWT) we can achieve a non-redundant N to N transformation. However, the DWT lacks the shift-invariance property, and in many applications, such as pattern recognition, non-shift invariance can lead to misleading results. One way to overcome the lack of shift-invariance property in the DWT is to remove the decimators and up-samplers in the transform and up-sample (insert zeros) the wavelet filters at each level in a dyadic fashion, that is by a factor $2^{(j-1)}$ at the j level. The resulting transform is overcomplete and has multiple names in the literature, such as the stationary wavelet transform (SWT) [18], ϵ-decimated wavelet transform [18], undecimated wavelet transform and "Algorithme á Trous" [19]. It is important to note that SWT is only defined for signals of length divisible by $2J$ where J is the maximum wavelet transform decomposition level.

In this paper we will refer to this transform as the Stationary wavelet transform; it is a shift-invariant, overcomplete transform that is able to provide shift invariance at the cost of overcompleteness. It has also been shown that a large class of processes can be captured by models based on stationary wavelets [18, 20].

3. FAULT DETECTION WITH STATIONARY WAVELET TRANSFORM

In this section we will present our novel method to detect faults on a three-phase low-voltage data with multi-scale stationary wavelet transform and sparse coding. The three-phase voltage signals have a sinusoidal structure with 120 degrees phase difference between the different phases. In the time domain, we can design the dictionary by dividing a period of the signal at each phase to multiple segments [6]. In the wavelet domain however, given that we have multiple scales, there are different ways of designing the dictionary. Similar to the time domain approach, we can divide the signal into different segments at each scale. Then we can either design multi-scale atoms where different scales are merged into one atom or we can have each segment of each scale as an atom. The dictionary design in the latter case can be expressed as:

$$D = [D_1 \quad D_2 \quad \dots \quad D_N], \quad (3)$$

where N is the number of wavelet scales. The three-phase voltage data we consider[1] are sampled at 4KHz, resulting in 80 samples per period. Due to speed limitations, we assume data vectors have length of 8 samples, therefore we only take 8 samples per phase for the length of each atom of the dictionary, resulting in 24 samples per atom. Given the number of samples, we take the stationary wavelet transform up to 2 levels, resulting in 4 scales. This process can also be done with dictionary learning methods such as K-SVD. Our empirical results show that, since the healthy voltage data do not evolve over time, dictionary with analytical design described above, performs better. The discussion above can be summarized as follows:

$$\min_S \frac{1}{2}||W_A Y - DS||_F^2 + \lambda||S||_1. \quad (4)$$

[1] Courtery of Kehui Ltd.

Here, W_A denotes the wavelet analysis operator, yielding $W_A Y$ to be the wavelet transform of the data vectors. We have changed all the vectors to their matrix format with capital letters to show that this method can be applied both online and in batch format. For the sake of simplicity, we omit W_A and assume the data vectors are already transformed using the Stationary wavelet transform. Our aim is to detect faults in the data and capture them in a separate vector. This is the underlying principle used for anomaly detection in [4] and [6]. With this approach in mind, we can write the following convex minimisation model for the fault detection:

$$\min_{S,E} \quad \frac{1}{2}||Y - DS - E||_F^2 + \lambda_1||S||_1 + \lambda_2||E||_1. \qquad (5)$$

Here, S is the sparse coding vector (or matrix for batch coding), E is the error term which absorbs anomalies and λ_1 and λ_2 are regularisation parameters for the sparse coding and error term respectively. The above convex minimisation problem can solved with the alternating direction method of multipliers (ADMM) [7]. The ADMM method, an augmented Lagrangian method, is a popular technique due to its simplicity and its fast convergence. To make the above minimisation problem ADMM compliant, we need to add an auxiliary term and its corresponding constraint as follows:

$$\begin{aligned} \min_{L,S,Z} \quad & \frac{1}{2}||Y - DS - E||_F^2 + \lambda_1||Z||_1 + \lambda_2||E||_1, \\ \text{subject to} \quad & Z = S. \end{aligned}$$

The augmented Lagrangian function of the above problem is:

$$\mathcal{L}(S, Z, E, U) = \frac{1}{2}||Y - DS - E||_F^2 + \lambda_1||Z||_1 + \lambda_2||E||_1 + U^T(Z - S) + \frac{\beta}{2}||Z - S||_F^2,$$

where U is the called the Lagrangian multiplier and β is a regularisation parameter. Minimizing with respect to S, Z and E yields the ADMM's variable updates for this problem:

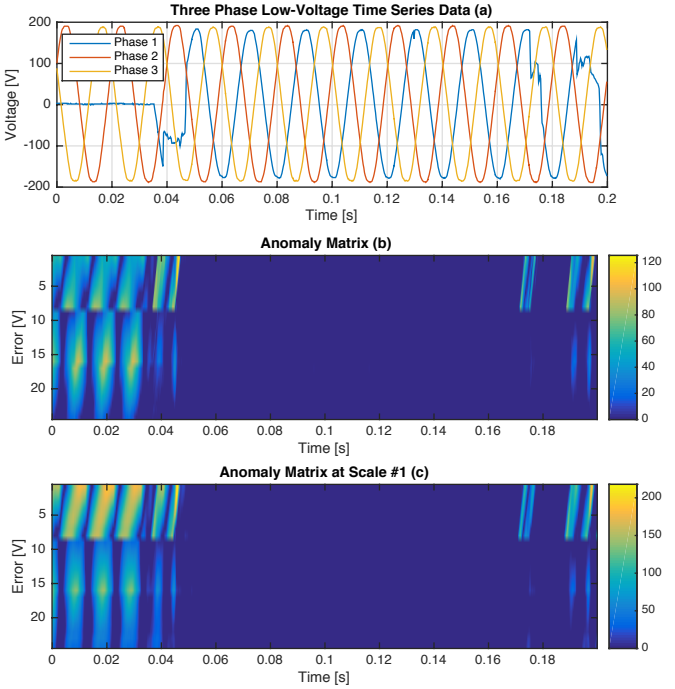
$$\begin{aligned} S^{K+1} &= \min_S \frac{1}{2}||Y - DS - E||_F^2 - U^T S + \frac{\beta}{2}||Z - S||_F^2 \\ Z^{K+1} &= \min_Z \lambda_1||Z||_1 + U^T Z + \frac{\beta}{2}||Z - S||_F^2 \\ E^{K+1} &= \min_E \frac{1}{2}||Y - DS - E||_F^2 + \lambda_2||E||_1 \\ U^{K+1} &= U^K + \beta(Z^{K+1} - S^{K+1}). \end{aligned}$$

By applying the minimisations, we achieve closed-form solutions for each variable. This separable structure gives the ability to implement the convex minimisation in a parallel framework. The fault detection algorithm is presented in Algorithm I. Here, $\mathcal{S}$ denotes element-wise soft-thresholding, $\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}$ are vectors of the regularisation parameters for all the wavelet scales and ϵ denotes the stopping criterion. Generally $\epsilon = 10^{-3}$ was a reasonable number, in terms of speed and accuracy, for most of our tests. Having presented the fault detection algorithm, we now illustrate how the detection performs compared to similar algorithms. Figure 1a shows a

Algorithm 1 ADMM steps for Fault Detection

1: **Initializations:** $S = 0$, $Z = 0$, $E = 0$; $U = 0$; $k = 0$, iter = 250.
2: **while** not converged & $k \leq$ iter **do:**
3: $S^{K+1} = (D^T D + \beta\mathbb{I})^{-1}(D^T(Y - E^K) + U^K + \beta Z^K)$
4: $Z^{K+1} = \mathcal{S}_{\frac{\boldsymbol{\lambda_1}}{\beta}}(S^{K+1} - \frac{1}{\beta}U^K)$
5: $E^{K+1} = \mathcal{S}_{\boldsymbol{\lambda_2}}(Y - DS^{K+1})$
6: $U^{K+1} = U^K + \beta(Z^{K+1} - S^{K+1})$
7: END LOOP if:
8: $\frac{||S^{K+1} - Z^K||_2^2}{||S^K||_2^2} < \epsilon$
9: $k = k + 1$

sample three-phase low-voltage time series data with multiple faults. We apply the minimisation algorithm described in Algorithm I to detect faults in the time series data. Figure 1b shows the anomaly matrix when dictionary is designed using the time domain samples. The regularisation and the ADMM parameters are $\lambda_1 = 5$, $\lambda_2 = 10$ and $\beta = 1$. The process can detect the faults at 0:0.04s, 0.17:0.18s and 0.19:0.2s. However, it can not detect the fault at 0.13s. Figures 1c up to 1f show the anomaly matrix at 4 different scales with stationary wavelet transform. The regularisation and the ADMM parameters are $\boldsymbol{\lambda_1} = [5, 5, 5, 5]$, $\boldsymbol{\lambda_2} = [15, 15, 3, 3]$ and $\beta = 1$. All these parameters were obtained by cross-validation. Generally, the coarse scales require higher values and the detail scales require smaller values for λ_2. As can be observed, Figures 1e and 1f, which are the detail scales of the transform, can detect the fault accurately at 0.13s. We can also observe from all the wavelet scales output, that our methodology has also correctly detected the healthy signals.



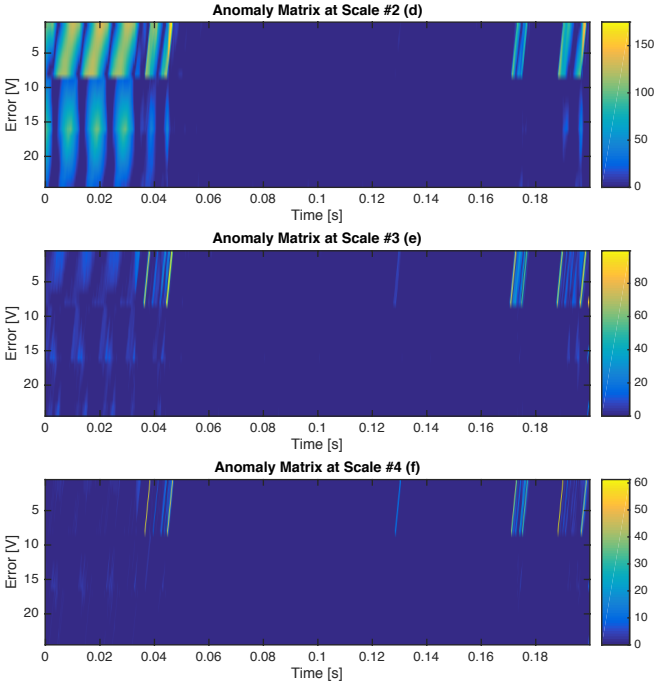


Fig. 1. Fault Detection in time and wavelet domain. (a) Three-phase low-voltage time series data with multiple faults. (b) Fault detector (anomaly) matrix using the time domain samples. (c:f) Fault detector (anomaly) matrix using stationary wavelet transform at scales 1,2,3 and 4 respectively.

4. FAULT CLASSIFICATION WITH STATIONARY WAVELET TRANSFORM

In the previous section we explained how stationary wavelet transform with convex optimization can be utilized to detect faults very accurately. The type of fault occurred is also of significant importance to the power distribution industry. In this section we will show how the same methodology can be used to detect and classify faults simultaneously. Fault types can typically be categorized by their phase, amplitude and frequency behaviours. The faults manifest as: a reduction in magnitude of one or two of the phases; or a syncing of two of the phases. As such, including the healthy class, there are therefore ten signal classes [6]. The approach we take is very similar to the previous section, except with two differences: first, the dictionary needs to also include the signatures of different faults, either designed analytically or learnt through a training set. Second, we can take benefit of the classification scheme and apply group-sparsity on the sparse coding vector. The updated minimisation model becomes:

$$\min_{S,E} \quad \frac{1}{2}||Y - DS - E||_F^2 + \lambda_1||S||_{2,1} + \lambda_2||E||_1. \qquad (6)$$

Here, $||S||_{2,1}$ is an $\ell_{2,1}$ norm which groups the components of S into $K = 10$ groups (classes) such that it imposes group-lasso penalty [21] on S: $||S||_{2,1} = \sum_{k=1}^{K} ||S_k||_2$. The sparsity is now on the level of the K classes and not the atoms of the dictionary. With the addition of $||S||_{2,1}$, Algorithm I, needs also to be updated. This results in a column-shrinkage

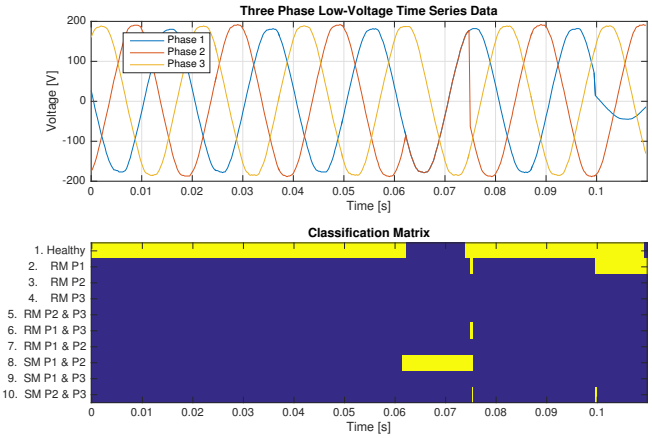


Fig. 2. Simultaneous fault detection and classification. (a) A three-phase low-voltage time series data with multiple faults. (b) The output of the classification where the results at each wavelet scale is normalized and combined to give a final classification matrix.

operator on Z. The update is therefore:

$$Z^{K+1}(:,i) = \begin{cases} P(:,i)\frac{||P(:,i)||_2 - \frac{\lambda_1}{\beta}}{||P(:,i)||_2} & \frac{\lambda_1}{\beta} < ||P(:,i)||_2 \\ 0 & \text{otherwise.} \end{cases}$$

Here, $P = (S^{K+1} - \frac{1}{\beta}U^K)$ and $P(:,i)$ is the i-th column of P. We now illustrate how the above proposed model performs for simultaneous fault detection and classification. Figure 2a shows a three-phase low-voltage time series data with two faults. The first fault occurs from around 0.06s where we can see that the phases of phase 1 and 2 are not 120 degrees apart and are identical. The second fault occurs at 0.1s and it is a reduced magnitude phase 1 fault. All the classification outputs from all wavelet scales are normalized and added together to achieve one final classification matrix, shown in Figure 2b. The regularisation and the ADMM parameters are $\lambda_1 = [30, 30, 30, 30]$, $\lambda_2 = [7, 7, 3, 3]$ and $\beta = 1$. It can be observed that the method has perfectly detected and correctly classified both faults.

5. CONCLUSION

In this paper we presented a novel method for anomaly detection using multi-scale stationary wavelet transform and sparse coding. We showed that our method can accurately detect anomalies in the data which are not picked up by the standard sparse error-coding schemes. We also presented an extension to the method to incorporate fault classification and showed that our method can simultaneously detect and classify faults. The faults we considered for the classification are faults that might not be entirely realistic. One possible future direction for this paper is to design a scheme that can retrieve extra information from the stationary wavelet domain in order to classify more complicated faults.

6. ACKNOWLEDGEMENT

We want to thank Innovations UK (IUK) for funding (via EPSRC and UCL) and also Kehui Ltd for providing us with three-phase low-voltage time series data.

7. REFERENCES

[1] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *Image Processing, IEEE Transactions on*, vol. 15, no. 12, pp. 3736–3745, 2006.

[2] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[3] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[4] A. Adler, M. Elad, Y. Hel-Or, and E. Rivlin, "Sparse coding with anomaly detection," *Journal of Signal Processing Systems*, vol. 79, no. 2, pp. 179–188, 2015.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 15, 2009.

[6] A. Kalaitzis and J. D. B. Nelson, "Online joint classification and anomaly detection via sparse coding," in *Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on*. IEEE, 2014, pp. 1–6.

[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[8] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397–3415, 1993.

[9] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*. IEEE, 1993, pp. 40–44.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[11] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[12] I. Tošić and P. Frossard, "Dictionary learning," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 27–38, 2011.

[13] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[14] I. Jolliffe, *Principal component analysis*, Wiley Online Library, 2002.

[15] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 12, pp. 1945–1959, 2005.

[16] K. Engan, S. O. Aase, and Hakon H. J., "Method of optimal directions for frame design," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*. IEEE, 1999, vol. 5, pp. 2443–2446.

[17] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *Signal Processing, IEEE Transactions on*, vol. 58, no. 3, pp. 1553–1564, 2010.

[18] G. P. Nason and B. W. Silverman, "The stationary wavelet transform and some statistical applications," *LECTURE NOTES IN STATISTICS-NEW YORK-SPRINGER VERLAG-*, pp. 281–281, 1995.

[19] S. Mallat, *A wavelet tour of signal processing*, Academic press, 1999.

[20] G. P. Nason, R. Von Sachs, and G. Kroisandt, "Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum," *Journal of the Royal Statistical Society, Series B*, vol. 62, pp. 271–292, 2000.

[21] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *arXiv preprint arXiv:1001.0736*, 2010.